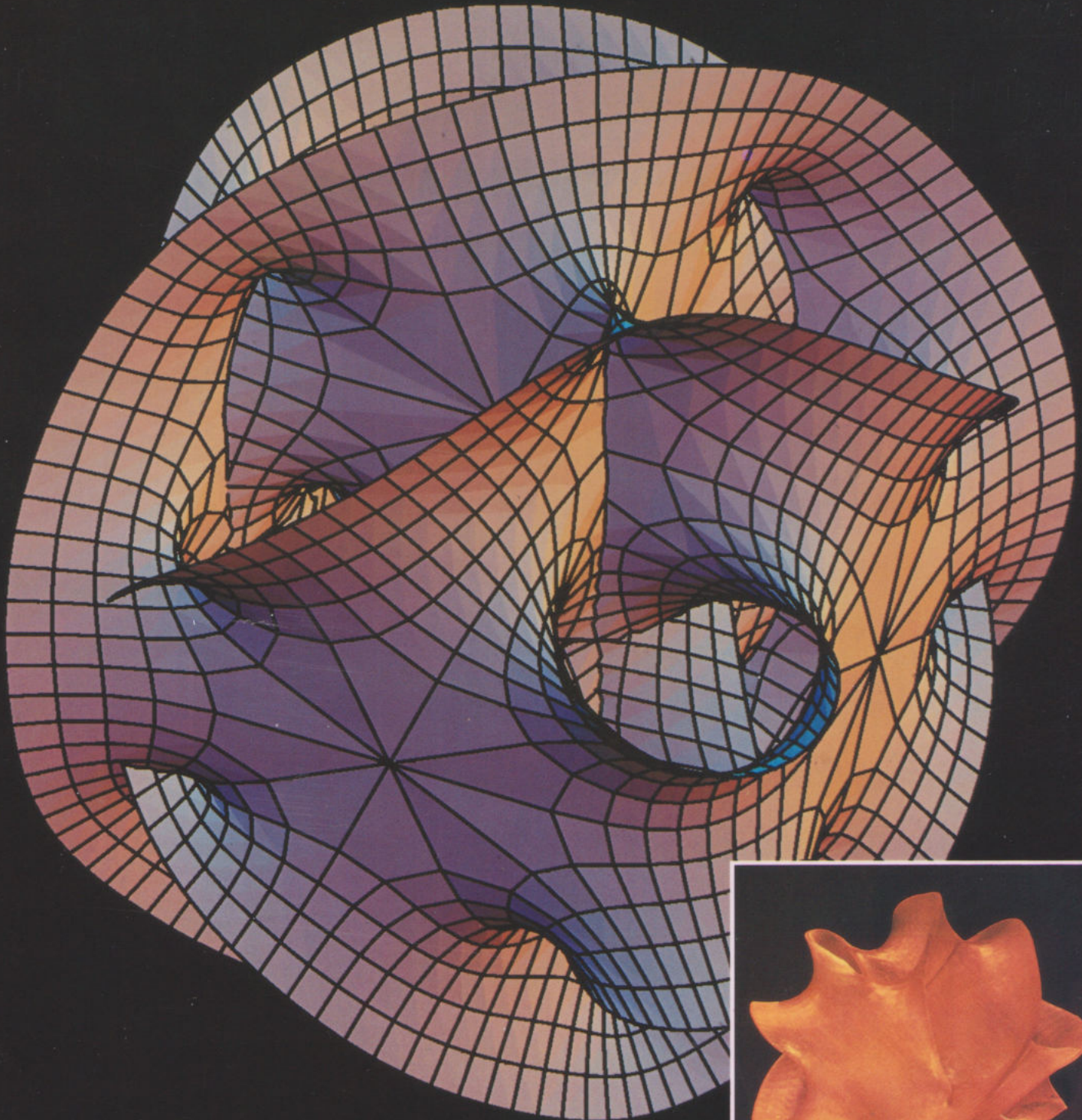


# What's Happening in the Mathematical Sciences

Volume 2



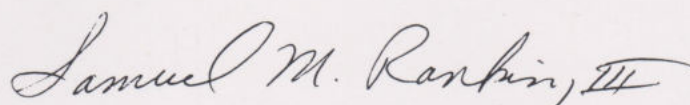
American Mathematical Society



## Introduction

Welcome to the 1994 issue of *What's Happening in the Mathematical Sciences*, a yearly publication of the American Mathematical Society, inaugurated in 1993. Volume 2 continues the theme of surveying some of the important developments in the mathematical sciences over the past year or so. One purpose of *What's Happening* is to convey that mathematics is a dynamic discipline, contributing to research and development in many areas of science as well as contributing significantly to the solving of some of the major problems facing society. In this issue you can read about a mathematically-based technology that produces real time continuous images of the heart, lungs, and other organs; results on key problems in the area of knot theory and how these results lead to insights in the study of DNA; recent findings in the theory of waves; and Fermat's Last Theorem.

*What's Happening in the Mathematical Sciences* is written in a style so that the general public can learn about the beauty and universality of mathematics. The American Mathematical Society hopes you enjoy it.



Samuel M. Rankin, III  
AMS Associate Executive Director

**Front Cover.** A collaboration between computer scientist Andrew Hanson at Indiana University and artist Stewart Dickson in Los Angeles has brought the Fermat equation  $x^n + y^n = z^n$  to life. The computer graphic shows a 3-dimensional projection of the complex Fermat surface  $u^5 + v^5 = 1$  (the exponent is indicated by the 5 grid lines that intersect at a point). Dickson has used a high-tech process called stereolithography to render the surface as a truly 3-dimensional sculpture.

**Back Cover.** New York-based sculptor Rhonda Roland Shearer combines elements of modern fractal geometry, expressed through plant forms, with classical Euclidean geometry in *The 5 Platonic Solids*: *Terra* (blue patina cube), *Ignis* (yellow ochre patina tetrahedron), *Aqua* (red patina dodecahedron), *Aer* (orange patina octahedron), and *Caelum* (viridian green patina icosahedron). (Photo courtesy of Lee Boltin. Copyright © 1992, by permission of Rhonda Roland Shearer.)

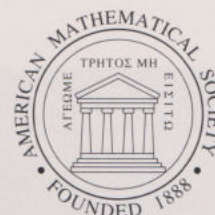
# What's Happening in the Mathematical Sciences

Volume 2

Written by Barry Cipra  
Edited by Paul Zorn

## Contents

<b>"A Truly Remarkable Proof"</b>	<b>3</b>
The announcement last year of a proof of Fermat's Last Theorem stunned the mathematical world. Andrew Wiles's proof, though currently incomplete, has nonetheless drawn rave reviews.	
<b>From Knot to Unknot</b>	<b>9</b>
What's the quickest way to untie a knot? Researchers have untangled a good part of the answer.	
<b>New Wave Mathematics</b>	<b>14</b>
Will compact waves cruise the information superhighways of the future? In theory, it's possible.	
<b>Mathematical Insights for Medical Imaging</b>	<b>19</b>
A team of mathematicians, computer scientists, and engineers has designed a new medical imaging technology based on the safe application of electric currents.	
<b>Parlez-vous Wavelets?</b>	<b>23</b>
Mathematicians and scientists are rapidly learning to speak a new language. The results are making a big splash.	
<b>Random Algorithms Leave Little to Chance</b>	<b>27</b>
Computer scientists will do anything to avoid bottlenecks and speed up computations. But gamble on the results? You bet!	
<b>Soap Solution</b>	<b>33</b>
Undergraduate students at a summer mathematics research program have found some slick answers to some old problems about the geometry of soap bubbles.	
<b>Straightening Out Nonlinear Codes</b>	<b>37</b>
A complicated class of error-correcting codes has suddenly gotten much easier to use.	
<b>Quite Easily Done</b>	<b>41</b>
A combinatorial problem, long thought to be difficult, has finally been solved—with surprising ease.	
<b>(Vector) Field of Dreams</b>	<b>47</b>
A clever construction "pulls the plug" on a 40-year-old conjecture about the topology of vector fields.	



ISBN 0-8218-8998-2  
ISSN 1065-9358

©1994 by the American Mathematical Society.  
All Rights Reserved.

Permission is granted to make and distribute verbatim copies of this publication or of individual items from this publication provided the copyright notice and this permission notice are preserved on all copies.

Permission is granted to copy and distribute modified versions of this publication or of individual items from this publication under the conditions for verbatim copying, provided that the entire resulting derived work is distributed under the terms of a permission notice identical to this one.

1991 *Mathematics Subject Classification*:  
Primary 00A06

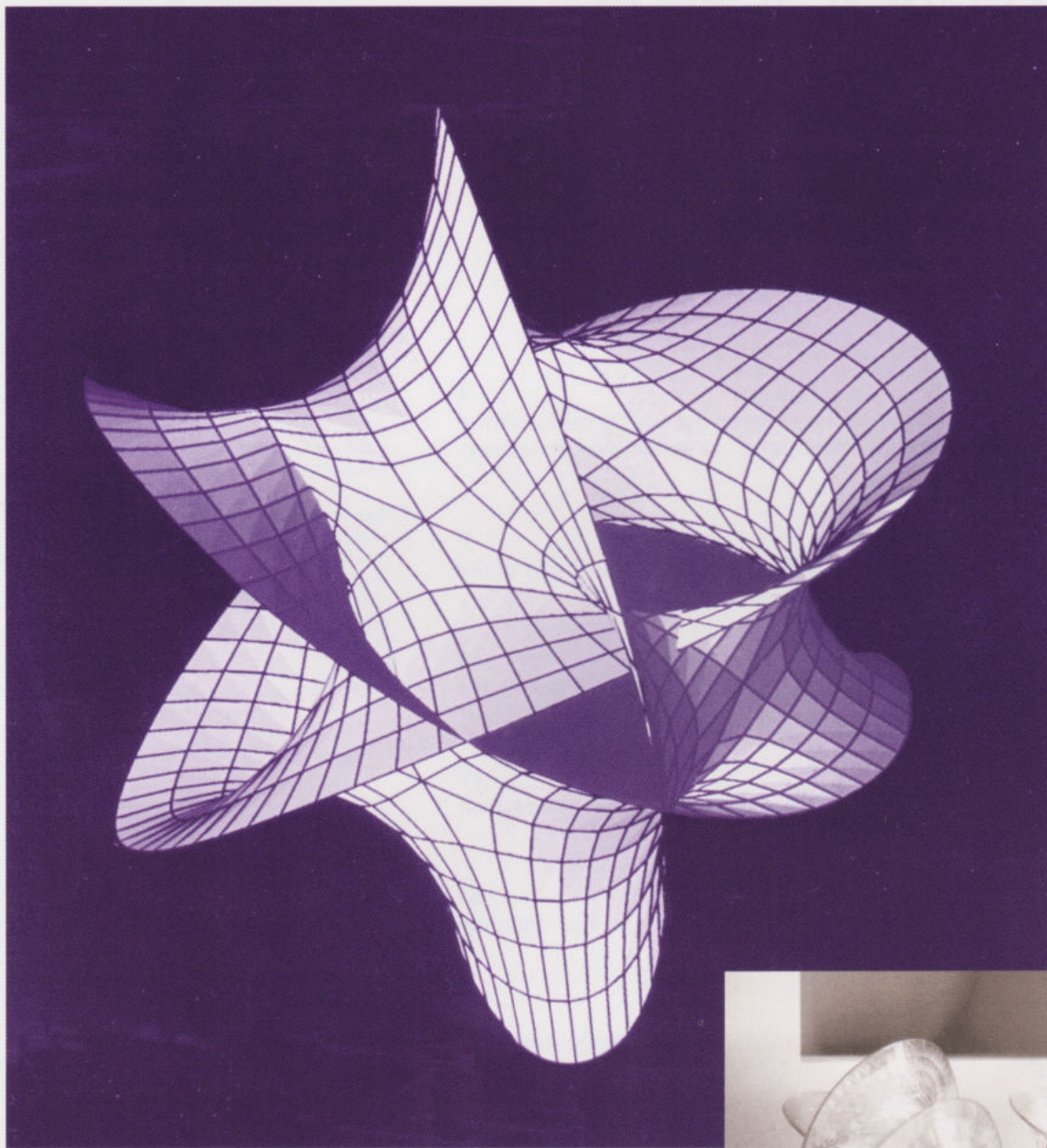
Printed in the United States of America.

This publication has been typeset using the  $\text{T}_{\text{E}}\text{X}$  typesetting system running on a Solbourne 5/502 Unix computer. Halftones were created from original photographs with Adobe Photoshop, and illustrations were redrawn using Adobe Illustrator on Macintosh Quadra computers. PostScript code was generated using **dvips** by Radical Eye Software.

Typeset on an Agfa/Compugraphic 9600 laser imagesetter at the American Mathematical Society. Printed at E. A. Johnson, East Providence, RI, on recycled paper.

10 9 8 7 6 5 4 3 2    05 04 03 02 01 00





**Figure 1.** A 3-dimensional projection of the complex Fermat surface  $u^3 + v^3 = 1$ , rendered with computer graphics (top) and through stereolithography as a plastic sculpture (bottom). (Figure courtesy of Stewart Dickson.)

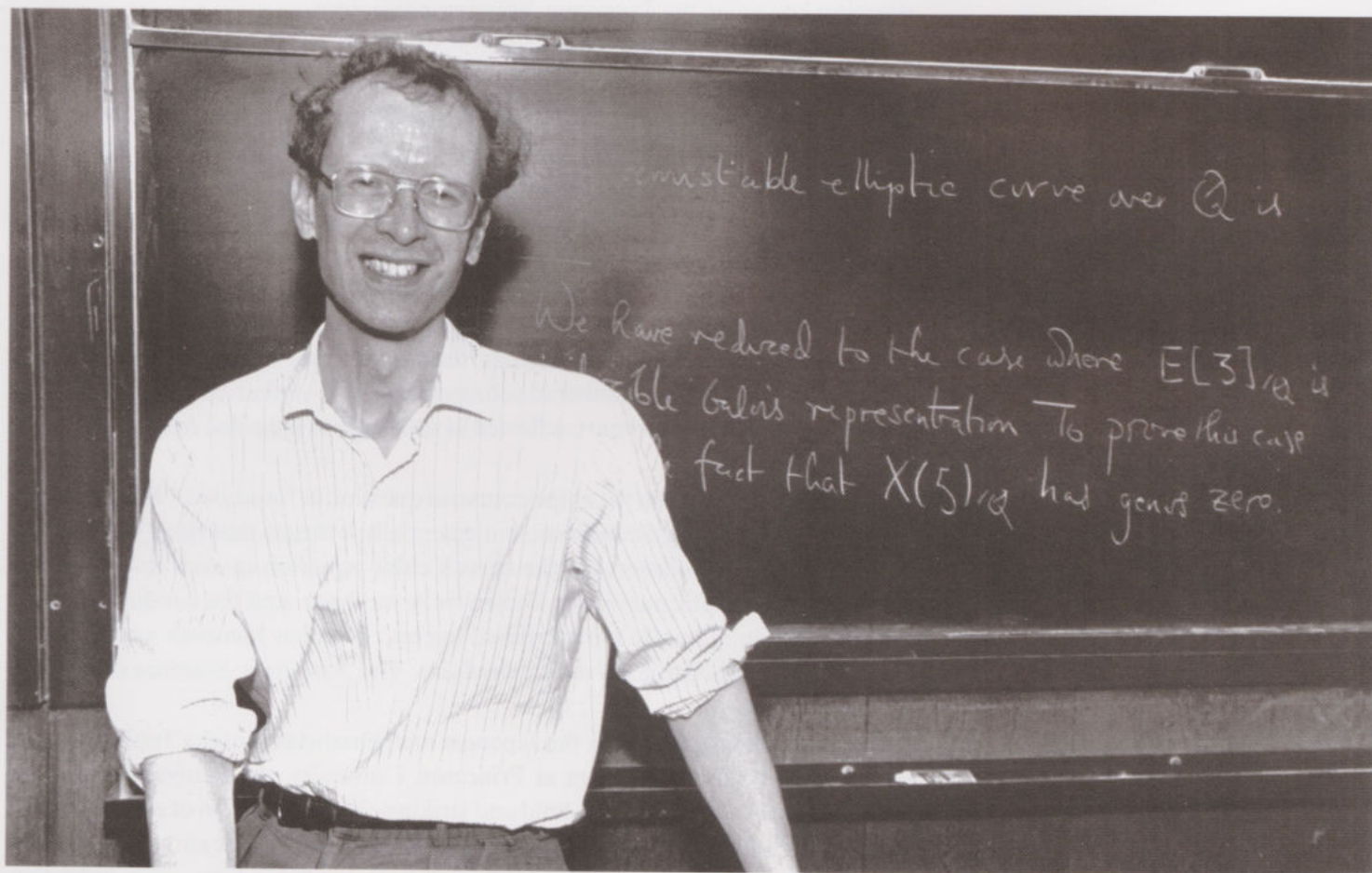


# “A Truly Remarkable Proof”

A torrent of electronic mail poured from Cambridge, England, on the morning of June 23, 1993. Mathematicians at a conference on number theory at the Isaac Newton Institute, a mathematical research center at the University of Cambridge, raced to tell their colleagues around the world some stunning news: Andrew Wiles, a number theorist at Princeton University, had just finished presenting a proof of Fermat's Last Theorem.

Wiles, it seemed, had solved mathematics' most famous open problem. Fermat's Last Theorem is a deceptively simple statement: The equation  $x^n + y^n = z^n$  has no solutions in positive integers  $x$ ,  $y$ , and  $z$  if the exponent  $n$  is greater than 2. The theorem was jotted down by the French mathematician Pierre de Fermat around 1637, in the margin of a math book, along with a tantalizing comment: “I have discovered a truly remarkable proof, which this margin is too small to contain.”

Countless mathematicians over the last 350 years have tried—and failed—to supply the missing proof. Prize money has even been offered for a solution. Curiously, by the usual standards of mathematics, the theorem itself is of little consequence: Unlike other famous unsolved problems in mathematics, Fermat's Last Theorem has no important corollaries. Rather, the problem's significance stems mainly from the theoretical machinery researchers have developed in trying



Andrew Wiles. (Photo courtesy of Denise Applewhite and Princeton University.)



---

**Number theorists would like to know whether *all* elliptic curves are modular. The Taniyama–Shimura conjecture says they are.**

---

to solve it. Indeed, most mathematicians long ago gave up working directly on Fermat's Last Theorem itself. Then Wiles dropped his bombshell in Cambridge.

The news lit up the mathematical world. It also grabbed the media's attention, as mathematical stories seldom do. Wiles's proof made the front page of the *New York Times*. It made *Time* and *Newsweek*. It made the NBC Nightly News ("Be still, my heart," said NBC's Tom Brokaw).

Experts who attended Wiles's lectures at the Newton Institute expressed confidence in the strategy of his proof, and amazement at the mathematical *tour de force* it represented. Still, mathematicians accept no proof as correct until it's been thoroughly checked—especially when the problem has the stature of Fermat's Last Theorem. And after the initial celebration had subsided and experts began meticulously poring over Wiles's 200-page manuscript, problems with the proof appeared. Most were minor, but one was not.

In early December, Wiles posted an e-mail message acknowledging a gap in the reasoning near the end of his proof. As this volume of *What's Happening* goes to press, the gap remains. Fermat's Last Theorem is still an open problem.

Yet number theorists continue to praise Wiles's work. "When people finally see this manuscript, they're just going to be bowled over completely," says an admiring Ken Ribet of the University of California at Berkeley. That's because Wiles's work, while aiming to prove Fermat's Last Theorem, advances number theory across a broad front. Indeed, the main focus of his work is not Fermat's Last Theorem itself, but one of the central problems in modern number theory, an assertion known as the Taniyama–Shimura conjecture.

To explain the Taniyama–Shimura conjecture and its relation to Fermat's Last Theorem requires a brief digression on the subject of elliptic curves. Roughly speaking, an elliptic curve is the set of solutions to a cubic equation in two variables. A typical equation, such as  $y^2 = x(x - 3)(x + 32)$ , sets the square of one variable equal to a cubic expression in the other. Number theorists are particularly interested in "rational points" on elliptic curves: solutions in which both  $x$  and  $y$  are rational numbers (see Figure 2).

One way to study the rational points on an elliptic curve is to look at the curve not in the ordinary system of real numbers, but in an infinite collection of *finite* number systems. In each finite system, the elliptic curve's cubic equation can be solved explicitly, and the number of solutions tallied. Number-theoretic properties of the original elliptic curve are reflected in solutions of the cubic equation in these finite systems.

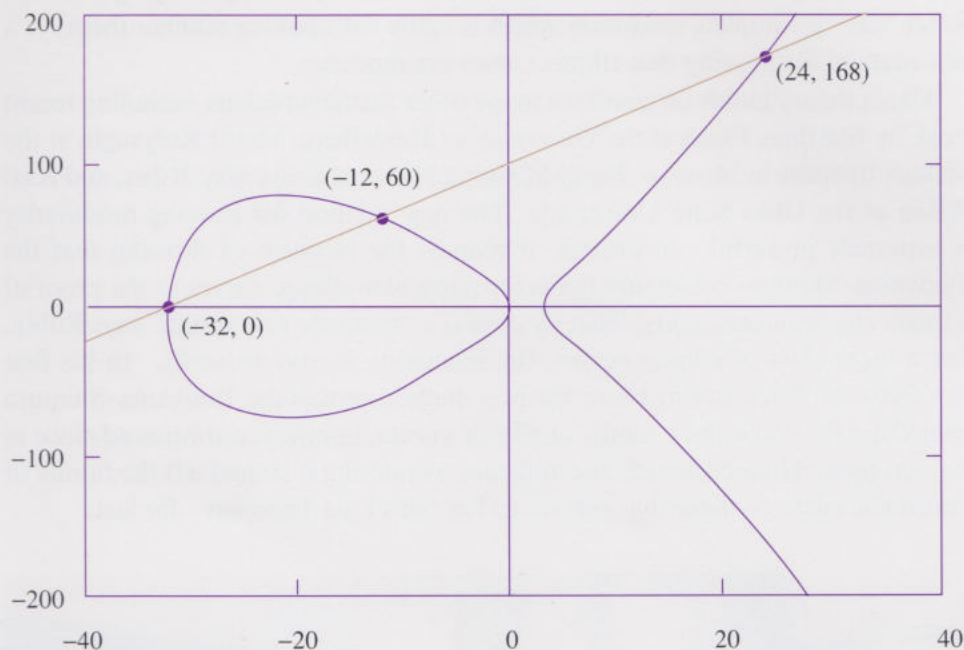
Things work best when the elliptic curve in question is "modular." Modularity is a complicated, technical condition, but essentially it means that there is a formula for the number of solutions of the curve's cubic equation in each finite number system. Many elliptic curves are known to be modular, and the condition can be checked computationally for individual curves. Number theorists would like to know whether *all* elliptic curves are modular. The Taniyama–Shimura conjecture says they are.

First formulated in 1955 by the Japanese mathematician Yutaka Taniyama, and later refined by Goro Shimura at Princeton University, the Taniyama–Shimura conjecture was—and still is—a bold and striking characterization of elliptic curves. In its full technical glory, the conjecture asserts that every elliptic curve is associated with a particular kind of function known as a modular form; this links two seemingly unrelated branches of number theory. The idea that there's a bridge



between elliptic curves and modular forms “really pervades lots of things that we do” in modern number theory, says Ribet. And unlike Fermat’s Last Theorem, the Taniyama–Shimura conjecture has a host of immediate consequences.

Fermat’s Last Theorem is one of them.



**Figure 2.** The elliptic curve  $y^2 = x(x-3)(x+32)$  has many rational points. A line connecting any two of them intersects at a third.

The connection between Fermat’s “simple” problem and the theory of elliptic curves came as a surprise when, in 1985, Gerhard Frey at the University of the Saarland in Saarbrücken, Germany, had the idea that any counterexample to Fermat’s Last Theorem could be used to construct a counterexample to the Taniyama–Shimura conjecture. Specifically, Frey proposed, if  $a^n + b^n = c^n$  for positive integers  $a$ ,  $b$ , and  $c$  and an exponent  $n$  greater than 2, then the elliptic curve with cubic equation  $y^2 = x(x - a^n)(x + b^n)$  cannot be modular.

Frey’s idea hinged on a technical result, which Jean-Pierre Serre at the Collège de France in Paris formulated as a precise conjecture. A year later, Ribet proved Serre’s conjecture. This established Fermat’s Last Theorem as a consequence of the Taniyama–Shimura conjecture.

Ribet’s result gave mathematicians a brand new way of thinking about Fermat’s Last Theorem and a new reason to work on the Taniyama–Shimura conjecture. Actually, it’s not necessary to establish the Taniyama–Shimura conjecture in full generality in order to deduce Fermat’s Last Theorem; it’s enough to prove it for a class known as semistable curves. This was the starting point for Wiles’s attack on the problem.

Wiles, who was already well known as an expert in the theory of elliptic curves, went to work full time on the Taniyama–Shimura conjecture. To avoid undue publicity he kept only one colleague at Princeton, Nicholas Katz, abreast of developments. Finally, in June, he asked to give three talks at the Newton Institute number theory conference. John Coates of Cambridge University, who was Wiles’s

**Ribet’s result gave mathematicians a brand new way of thinking about Fermat’s Last Theorem and a new reason to work on the Taniyama–Shimura conjecture.**



---

**In his third lecture, Wiles announced his major result: The Taniyama–Shimura conjecture is true for semistable elliptic curves.**

---

thesis advisor at Cambridge in the mid-1970s, scheduled him to speak on Monday, Tuesday, and Wednesday, June 21–23, 1993.

The audience could tell just from the title of his lectures—“Elliptic curves, modular forms, and Galois representations”—that Wiles had important news to impart, perhaps pertaining to Fermat’s Last Theorem. (All three items in Wiles’s title are key ingredients in Ribet’s 1986 result). What Wiles began laying out, says Ribet, was “a complete revelation which is really still shaking number theory”: a new method for proving that elliptic curves are modular.

Wiles’s theory builds on results of many other mathematicians, including recent work by Matthias Flach at the University of Heidelberg, Victor Kolyvagin at the Steklov Institute in Moscow, Barry Mazur at Harvard University, Ribet, and Karl Rubin at the Ohio State University. The new method for proving modularity is extremely powerful. In essence, it reduces the problem of showing that the Taniyama–Shimura conjecture holds for particular elliptic curves to the proof of a single algebraic inequality. That by itself is a “fantastic new result,” says Rubin. For a large class of elliptic curves, the inequality is easy to verify. In his first two lectures, Wiles outlined how the new method proves the Taniyama–Shimura conjecture for one infinite family of elliptic curves, another enormous advance in its own right. His lectures left the audience wondering if he had left the family of semistable curves—those that pertain to Fermat’s Last Theorem—for last.



*Left to right: John Coates, Andrew Wiles, Ken Ribet, and Karl Rubin at the Isaac Newton Institute in Cambridge, England, after Wiles’s historic talk. (Photo courtesy of Ken Ribet.)*

He had. In his third lecture, Wiles announced his major result: The Taniyama–Shimura conjecture is true for semistable elliptic curves. Almost as an afterthought, he noted the long-awaited corollary: Fermat’s Last Theorem. It took a moment for the announcement to sink in. Then the audience burst into applause.

“The logic of his argument is utterly compelling,” Ribet said at the time. Other number theorists agreed that Wiles had cleared many of the technical hurdles on the way to a proof of the Taniyama–Shimura conjecture and had set a new agenda for the theory of elliptic curves. However, the review process has revealed a gap near the end of the proof: The calculations that verify the crucial inequality, which are easy in some cases, turn out to be not so easy for the class of semistable curves.



Experts believe that Wiles's basic strategy for the calculations is sound, even if the details don't yet fit together.

It's not unusual for a long, complicated mathematical proof to contain an error. Wiles's colleagues are quick to point out (it's not even unusual for a *short* proof to be mistaken). Nobody knows how long it will take to fill the gap. Still, says Rubin, "it's hard to believe that the proof of Fermat's Last Theorem is not closer."

**Nobody knows how long it will take to fill the gap.**

### Fermat's Last Theorem is True (for Exponents up to 4,000,000)

Fermat's Last Theorem states that the equation  $x^n + y^n = z^n$  has no solutions in positive integers  $x$ ,  $y$ , and  $z$  if the exponent  $n$  is greater than 2. Could the theorem be true for some exponents and false for others? Mathematicians have made much progress in the last 350 years in showing that if any counterexamples exist, the numbers involved are colossal.

Although he never found room in the margin or anywhere else for a general proof, Fermat did write down a proof of his famous theorem for the special case  $n = 4$ . Over a hundred years later, the Swiss mathematician Leonhard Euler dispatched the case  $n = 3$ . In the 1820s and 1830s, the theorem was proved for exponents 5 and 7. (It's enough to prove Fermat's Last Theorem for *prime* exponents. For example, if  $x = a$ ,  $y = b$ , and  $z = c$  solve the equation  $x^6 + y^6 = z^6$ , then  $x = a^2$ ,  $y = b^2$ , and  $z = c^2$  solve  $x^3 + y^3 = z^3$ .)

In the 1840s, the theory took a giant leap forward. By introducing some potent new ideas, Ernst Eduard Kummer was able to prove Fermat's Last Theorem for all prime exponents up to 100, with the exception of three "irregular" primes. In Kummer's theory, primes are classified as either regular or irregular. Fermat's Last Theorem, the theory says, is true for all regular primes. Regular primes are believed to be more common than irregular ones, constituting roughly 60% of all primes. Ironically, though, while it's known that there are infinitely many irregular primes, the same statement (while undoubtedly true) has never been proved for regular primes.

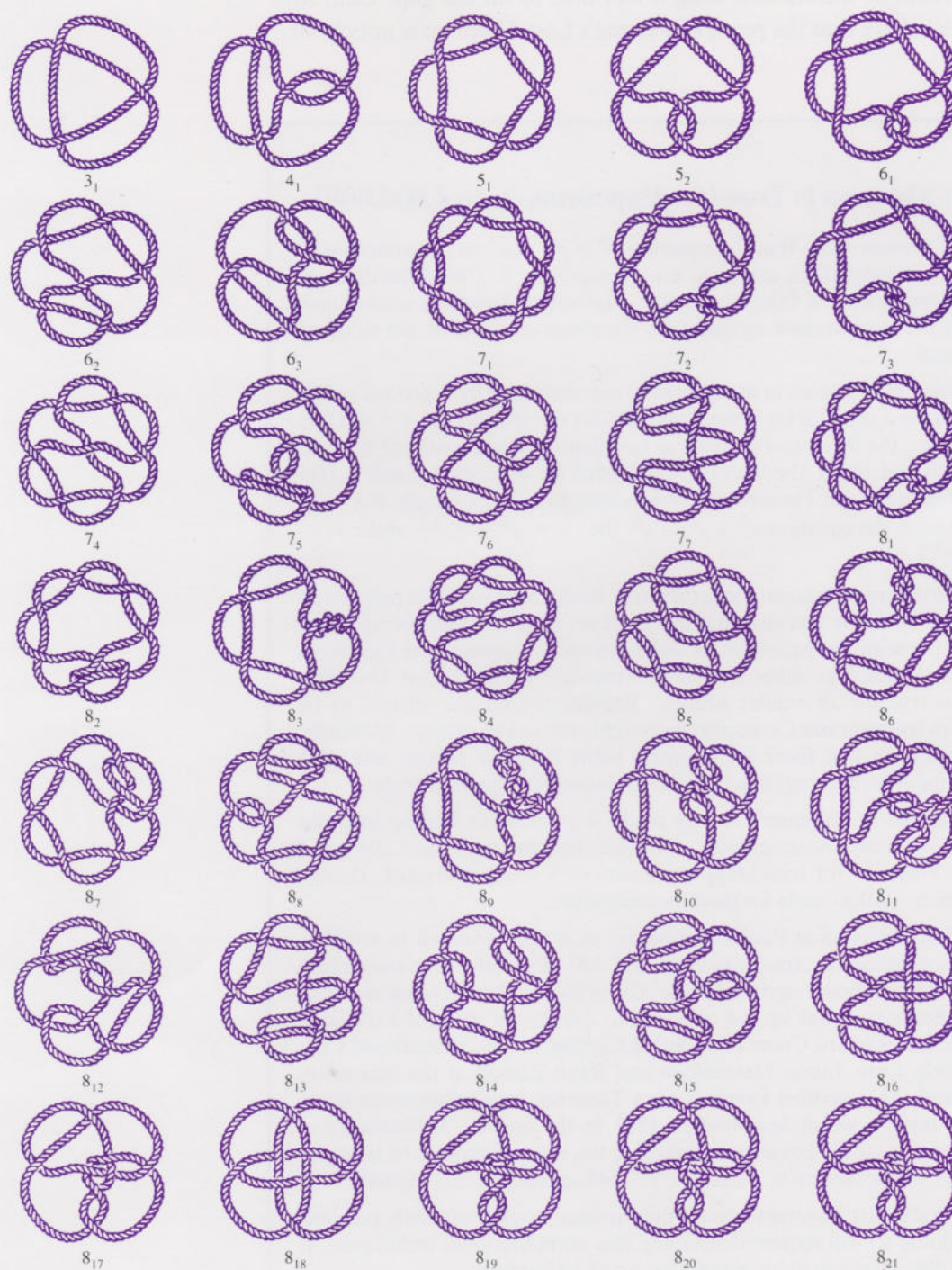
Later improvements in Kummer's theory made it possible to handle irregular primes separately, on a case-by-case basis. In effect, the theory reduces the proof of Fermat's Last Theorem for individual exponents to a straightforward, though lengthy, computation—tailor-made for modern computers.

In the 1970s, Sam Wagstaff at Purdue University used this approach to establish Fermat's Last Theorem for all exponents up to 125,000. Recently, four researchers have pushed the computational approach into the millions. Using refinements of Kummer's basic theory to speed up the calculation, Joe Buhler at Reed College in Portland, Oregon, and Richard Crandall at NeXT Computer Inc., in Redwood City, California, with help from Tauno Metsänkylä and Reijo Ernvall at the University of Turku in Finland, have verified Fermat's Last Theorem for all exponents up to 4 million. Their results, which appeared in 1993 in the journal *Mathematics of Computation*, also support the conjecture regarding the ratio of regular to irregular primes: Out of 283,145 primes up to 4 million, 171,548, or 60.59%, are regular.

Extending Fermat's Last Theorem beyond the 4 million mark is certainly possible, says Buhler, but doing so will require developing new computational techniques. If Wiles succeeds in filling the gap in his proof, that won't be necessary.



# Knots to 8 Crossings



David Broman and Charlie Gunn

**Figure 1.** Reprinted with permission from Supplement to Not Knot by David Epstein and Charlie Gunn, published by A K Peters, Ltd.



# From Knot To Unknot

Alexander the Great didn't mess around. As legend has it, the Macedonian king decided to try his luck with the fabled Gordian knot, a tough length of cornel bark wrapped tightly around the pole of an ox cart. It was said that the person who succeeded in untying this knot was destined to rule the world (meaning, at the time, Persia). A man of action rather than dexterity or patience, Big Al unsheathed his sword—and the rest, as they say, is history.

Modern mathematicians are also drawn to the problem of undoing knots, although their motives—and their techniques—are quite different from Alexander's. In the last few years, researchers using two different approaches have come to understand better just what it takes to unknot a knot.

A mathematical knot is basically just a closed curve that winds through 3-dimensional space, like an electrical extension cord that's been tangled up and then plugged into itself. The theory of these meandering curves has taken off in the last decade. "Knot theory for a long time was a little backwater of topology," notes Joan Birman, an expert in the subject at Columbia University. "It's now been recognized as a very deep phenomenon in many areas of mathematics." And it's not just mathematics where knot theory is playing a larger role; molecular biologists, for example, are using it to help untangle some of the geometric secrets of DNA (see box page 13).

One key problem in knot theory is to decide whether one knot can be deformed into another—in particular, to tell whether a given knot really isn't knotted at all. That may sound like a straightforward, even trivial, problem. But it's really as difficult to deal with as a snarled-up fishing line. The main difficulty is that there are infinitely many ways to deform any knot, and they all must be ruled out in order to show that two knots are indeed different.

Because it's hard to draw truly 3-dimensional pictures, knots are commonly represented by projections onto a plane. Such a picture, called a "knot diagram," can be thought of as tracing the path of a tangled extension cord that's been dropped onto the floor; places where the curve is broken are called crossings. The number of crossings depends on how the cord has been dropped, and can be quite large—but every knot has a diagram with a minimal number of crossings. Knot theorists have constructed elaborate tables of knots arranged according to this number (see Figure 1). These tables were begun in the 1890s by the British mathematician P. G. Tait, who was inspired by Lord Kelvin's theory that atoms were "knotted vortices" in the ether. (Kelvin's idea did not survive, but surprisingly, knot theory has re-emerged in physics, this time in an area known as quantum field theory.)

In the 1920s, the German mathematician Kurt Reidemeister showed that any deformation of a knot can be achieved by a sequence consisting of three types of moves (see Figure 2). This gives a combinatorial flavor to the topological problem of classifying knots, but it does not automatically solve the problem, because there are no set rules that specify the order in which the moves should be applied. For example, you might think that if a knot can be deformed into the "unknot" (the knot theorist's word for the circle), the deformation could be done without ever increasing the number of crossings. That's not true: For some diagrams, the crossing number must go up before it can come down (see Figure 3).

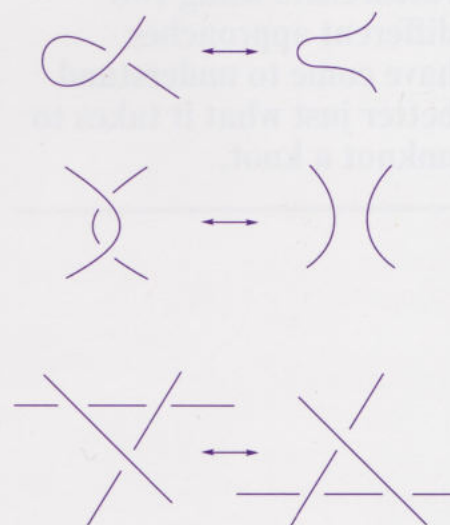


Figure 2. The three types of Reidemeister moves.

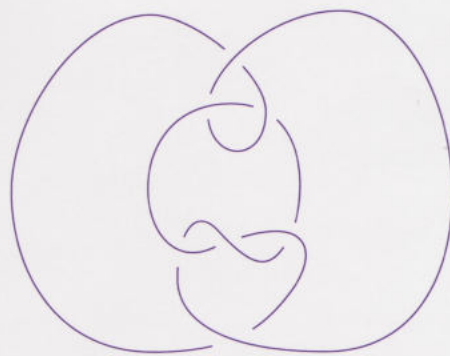


Figure 3. A "nasty" unknot that can only be unknotted by first increasing the number of crossings.

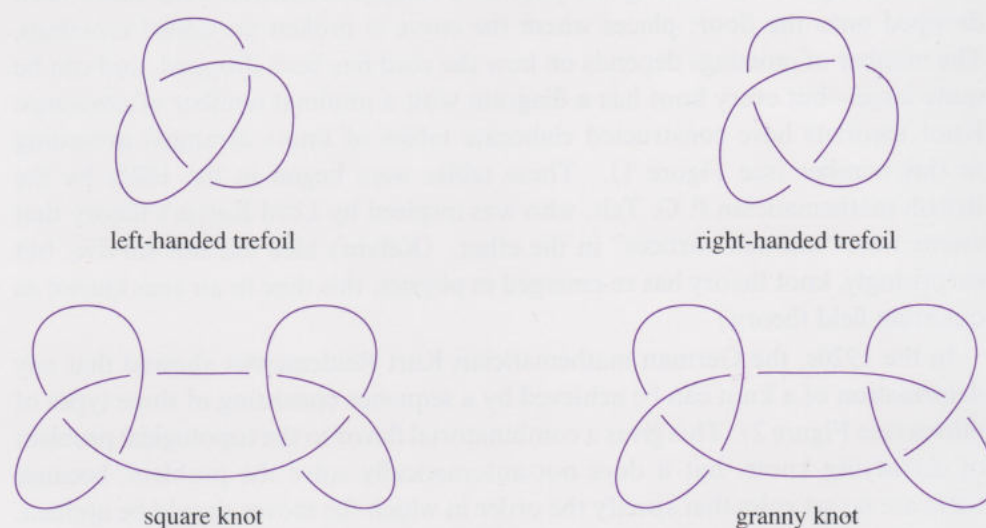


**In the last few years, researchers using two different approaches have come to understand better just what it takes to unknot a knot.**

So how *do* mathematicians decide whether two knots are different? Knot theorists' favorite approach has been to compute "invariants": numerical or algebraic expressions assigned to a knot that don't change when the knot is deformed. One of the earliest invariants, also dating back to the 1920s, is known as the Alexander polynomial (named after the American mathematician John Alexander, not Alexander the Great). Though defined topologically, the Alexander polynomial can be derived from the pattern of under- and over-crossings in a knot diagram. Most important, if two knots have different Alexander polynomials, then they are necessarily different knots. For example, the trefoil knot, whose polynomial is  $x^2 - x + 1$ , differs from the square knot, whose polynomial is  $(x^2 - x + 1)^2$ —and both differ from the unknot, whose polynomial is the constant 1 (see Figure 4). However, different knots need not have different Alexander polynomials. The granny knot, for example, has the same polynomial as the square knot. Likewise, the right- and left-handed trefoil knots share the same polynomial even though it's impossible to deform one into the other.

For a long time, though, the Alexander polynomial was one of the few tools topologists had for telling knots apart. Then in 1984, Vaughan Jones, a mathematician at the University of California at Berkeley, discovered a new polynomial invariant. Jones's polynomial turned out to be more powerful than Alexander's at distinguishing different knots. It also revealed startling new connections between knot theory and mathematical physics. More recently, Viktor Vassiliev at the Independent University of Moscow has introduced a whole new class of invariants based not on the topology of individual knots but on the structure of the space of *all* closed curves, even those that pass through themselves (such curves are viewed as degenerate, or "singular," knots). Birman and Xiao-Song Lin at Columbia University have found deep connections between Vassiliev's invariants and the Jones polynomial.

The Alexander and Jones polynomials are easy to compute, but they don't refer to anything that can be seen geometrically in a knot diagram. The minimal crossing number for a knot, on the other hand, refers explicitly to something that



**Figure 4.** The square knot is formed by joining a right- and left-handed trefoil; the granny knot is formed by joining two right-handed (or two left-handed) trefoils.



can be seen. So does the “unknotting number,” which is the least number of times you need to “cheat” by passing a knot through itself in order to untie it. But these invariants can be hard to compute.

Theorists generally compute a knot’s minimal crossing number by a process of elimination: First they find a diagram that seems to have the fewest crossings; then they show that the knot is different from every knot with fewer crossings, typically by comparing Alexander or Jones polynomials. The second step, however, requires a complete list of knots with smaller crossing numbers. So far that list is complete only up to crossing number 14.

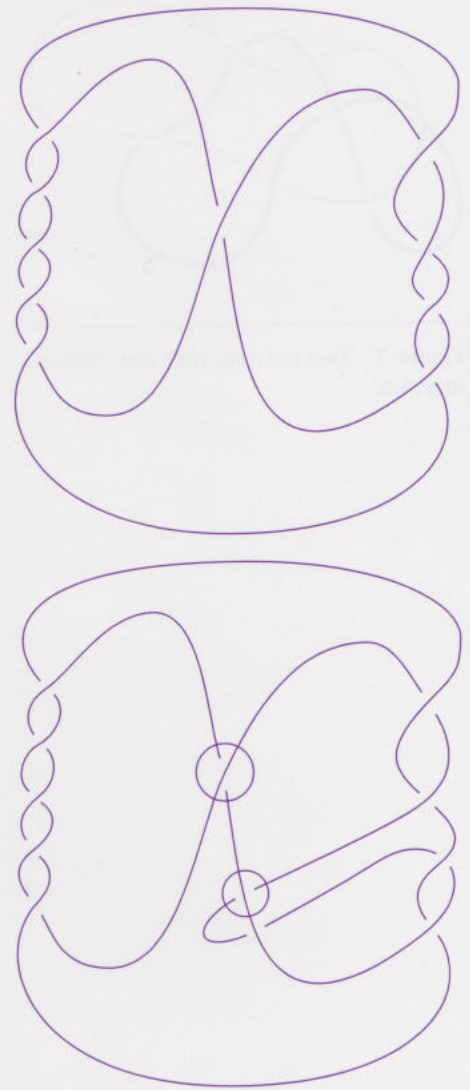
The unknotting number is even harder to compute. You might think you could compute it by taking any diagram and finding the smallest combination of crossings which, if the knot is passed through itself at those points (so that underpasses become overpasses and vice versa), will untie the knot. Unfortunately, that doesn’t necessarily give the right answer—it only puts an upper bound on the unknotting number. For example, Figure 5 (top) shows a knot diagram—in fact, one with a minimal number of crossings—which cannot be untied by changing fewer than three crossings. But Figure 5 (bottom) shows the same knot in a diagram that can now be untied with just two cheats. In other words, to find the fewest crossings to change, it may be necessary to take a simple looking diagram and redraw it to look more complicated; and there seems to be no bound on how much more complicated a knot diagram may need to look before it exhibits the correct unknotting number.

Until the recent breakthrough, theorists had no general method for computing unknotting numbers, except when the value happens to be 1 (if a knot can be untied with a single cheat, then its unknotting number must be either 1 or 0, so one need only check whether the knot was already unknotted). But researchers have recently proved results that allow knot theorists to compute unknotting numbers exactly for many more knots, and obtain useful lower bounds on unknotting numbers for *all* knots.

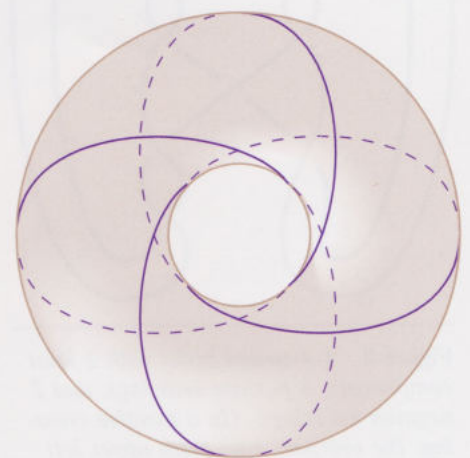
Working on problems in 4-dimensional topology, Peter Kronheimer at Oxford University and Tomasz Mrowka at the California Institute of Technology have proved a 40-year-old conjecture, due to John Milnor, about unknotting numbers for a special class of knots. Milnor’s conjecture specifies the unknotting number for all “torus” knots. These knots come from curves that are drawn on a torus, which is what mathematicians call a donut. To tie a  $(p, q)$ -torus knot, wrap a string  $p$  times through the hole in a donut, stretching the string so it goes  $q$  times around the donut itself before you tie the ends of the string together; then eat the donut (see Figure 6).

Milnor, who was also mainly interested in 4-dimensional topology, conjectured that the unknotting number for such a knot is always  $(p-1)(q-1)/2$  ( $p$  and  $q$  can’t both be even; in fact, in order for the knot to be drawn on the torus without intersecting itself,  $p$  and  $q$  can’t have any common divisor greater than 1). For example, Milnor’s conjecture says that the  $(101, 3)$ -torus knot has unknotting number 100. Given the difficulty knot theorists have had computing the unknotting number when its value is greater than 1, Milnor’s conjecture seems almost miraculous.

How, you might wonder, does 4-dimensional topology get mixed into the theory of knots? The trick is to view the deformation of a knot as occurring in time, which adds a fourth dimension to space. “If we think of a space–time picture of what is going on, the moving curve in space sweeps out a 2-dimensional surface in space–time,” Kronheimer explains. Topologically, the 2-dimensional surface is



**Figure 5.** The diagram at top cannot be unknotted with fewer than three changes of crossings, but the modified diagram below can be unknotted with only two (indicated by circles).

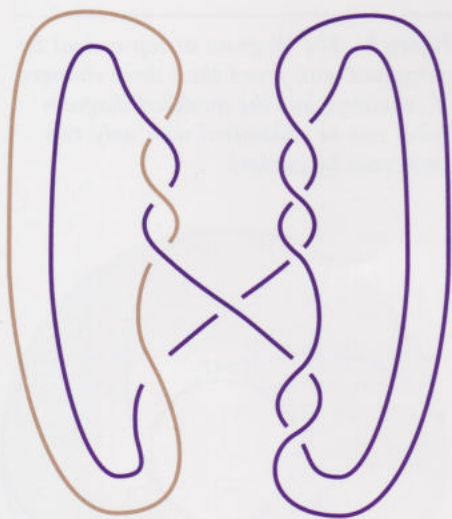


**Figure 6.** A  $(4,3)$ -torus knot.





**Figure 7.** Two unknots that are “linked” together.



**Figure 8.** A 4-strand braid with 2 knot components, 8 positive crossings, and 2 negative crossings. (In a positive crossing, the overpass goes from upper left to lower right; in a negative crossing, it goes from upper right to lower left.)

like a cylinder, he adds, but “because the original curve is knotted, the cylinder sits in space–time in a rather complicated way.”

If the deformation includes a cheat, then the space–time surface intersects itself at that point. Kronheimer and Mrowka were studying the general theory of self-intersecting, or “immersed,” surfaces in 4-dimensional space. Their research is based on far-reaching ideas introduced in the early 1980s by Simon Donaldson at Oxford University, who borrowed techniques from theoretical physics to analyze the structure of 4-dimensional spaces. The result for knots came as a kind of 3-d bonus. “We weren’t really aiming at Milnor’s [conjecture],” Kronheimer says.

More recently, Lee Rudolph at Clark University in Worcester, Massachusetts, has shown that Kronheimer and Mrowka’s results also prove a generalization of Milnor’s conjecture due to Daniel Bennequin at the University of Strasbourg in France, which provides a lower bound on the unknotting number for all knots, and all “links” as well. (A link is simply a set of knots that are tangled together, as in Figure 7.) Bennequin’s conjecture requires drawing the knot (or link) diagram in a particular configuration known as a braid and distinguishes between “positive” and “negative” crossings (see Figure 8). If a braid with  $M$  strands and  $R$  components (which is 1 for a knot, and greater than 1 for a link) has  $P$  positive and  $N$  negative crossings, then Bennequin’s conjecture asserts that the unknotting number  $U$  satisfies the inequalities  $|P - N| \leq 2U + M - R \leq P + N$ . If all crossings have the same sign (say positive), then Bennequin’s conjecture gives an exact value for the unknotting number. In particular, it turns out that a  $(p, q)$ -torus knot can be drawn as a braid with  $p$  strands and  $(p - 1)q$  positive crossings, so Milnor’s formula falls out of Bennequin’s conjecture. (Actually, only the lower bound in Bennequin’s conjecture required proof; the upper bound was proved by Michel Boileau at the University of Toulouse and Claude Weber at the University of Geneva in 1983, shortly after the conjecture appeared.)

As if one proof weren’t enough, William Menasco, a knot theorist at the State University of New York at Buffalo, has also proved Bennequin’s conjecture, using completely different methods. (This parallels the legend of Alexander. According to some accounts, the Great one didn’t draw his sword, but instead removed the pole on which the Gordian knot was tied, leaving the knot to fall apart of its own accord.) Working independently at about the same time as Kronheimer and Mrowka, Menasco actually proved a stronger version of Bennequin’s conjecture, one that looks more closely at the distinction between positive and negative crossings in a knot or link.

In Menasco’s theorem, the unknotting number is replaced with positive and negative variants. The positive unknotting number,  $U_+$ , is defined as the minimal number of positive crossings that must be changed to negative ones in order to untie a knot, regardless of how many negative crossings must be changed to positive. (The negative unknotting number,  $U_-$ , is defined similarly.) Menasco showed that  $U_+$  satisfies the inequality  $P - N \leq 2U_+ + M - R$ , provided that  $P \geq N$ . (A similar inequality holds for  $U_-$  if  $P \leq N$ ). Since the original unknotting number  $U$  is never less than  $U_+$  or  $U_-$ , Menasco’s inequalities together imply Bennequin’s conjecture.

Menasco’s proof is strictly 3-dimensional. Like Kronheimer and Mrowka’s proof, it is based on a careful study of immersed surfaces, but in this case the surfaces are deformed disks bounded by knots, all situated in ordinary 3-dimensional space. The proof is “very geometric” and involves “a lot of picture drawing,” Menasco says, adding that his approach uses “low-tech mathematics” compared



to the methods employed by Kronheimer and Mrowka. Birman, who has collaborated with Menasco on an extensive study of links and braids, disagrees. "The proof that he found is very hard," she says. "Some of the things that he did are extremely difficult to visualize. His ability to visualize 3-dimensional geometry is rather extraordinary."

Birman is enthusiastic about the new results on the unknotting number. "It's the beginning of a real theory of this mysterious number," she notes. She is also pleased that there are two proofs. "It's quite wonderful that two such widely different techniques could lead to the same result," she says. "I think it's evidence of the unity of mathematics."

But that's what knots are good for: Tying things together.

### The Knotted Helix

Mathematicians aren't the only ones excited by the latest results in knot theory. Molecular biologists, too, are eager to get in on the action.

"For me it's great," says Sylvia Spengler, a molecular biologist at the University of California at Berkeley. "It gives me insight on how frequently an enzyme had to act."

Spengler is one of a growing group of researchers applying theorems from topology to the chemistry of life. That may seem like a stretch—but stretching is what topology is all about. Biologists have long known that DNA is not only wound in a double helix, but also tightly coiled inside the nucleus of the cell. But only recently have researchers begun to understand the details of what they call supercoiling.

Supercoiling is found, for example, in circular DNA, a form of the macromolecule that occurs in bacteria and yeast. The flexible molecule need not look like a geometric circle, though; it may even be knotted. Knotting—and unknotting—is caused by enzymes called topoisomerases. These enzymes cut the strand of DNA at one point, pass another part of the strand through the gap, and then reseal the cut—exactly what's called for in the unknotting number theorem.

De Witt Summers, a knot theorist at Florida State University who collaborates with Spengler and others on topological aspects of molecular biology, points out that the unknotting number is a lower bound for the number of times the topoisomerase has to act. "If you have really complicated products that have a large unknotting number, it's going to take the enzyme a while to produce those," he explains.

**According to some accounts, Alexander the Great didn't draw his sword, but instead removed the pole on which the Gordian knot was tied, leaving the knot to fall apart of its own accord.**

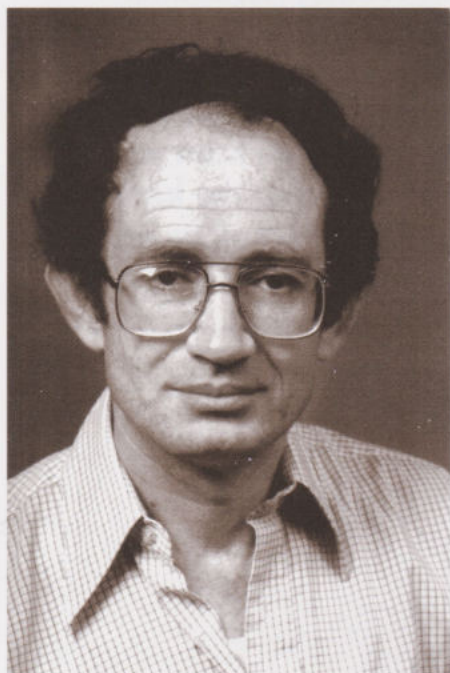


**Figure 9.** Three strands of knotted circular DNA. (Photo courtesy of Sylvia Spengler, University of California, Berkeley, and Frank Dean, Rockefeller University.)



# New Wave Mathematics

---



Philip Rosenau.

**J**ohn Scott Russell knew he was on to something one August day in 1834, when he chased a peculiar “heap of water” down the Edinburgh–Glasgow canal. When a boat being pulled by horses suddenly stopped, the wave “rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth, and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed.” The wave, he noticed, was approximately 30 feet long and a foot and a half high, and traveled at 8 or 9 miles per hour. Russell followed it on horseback for a couple of miles until it finally disappeared. “Such,” he later wrote, “was my first chance interview with the singular and beautiful phenomenon which I have called the Wave of Translation.”

At the time, Russell’s observation was considered an anomaly; it was even greeted with disbelief. These days, the theory of solitary waves is a well developed subject, with close ties to mathematical physics. But even so, there are still surprises and potential applications waiting in the wings. One surprise surfaced recently when Philip Rosenau, a theorist at the Technion in Haifa, Israel, discovered a class of waves so solitary that two of them can move along within a hair’s breadth of each other yet remain blissfully unaware of each other’s existence. Rosenau and Mac Hyman, a mathematician in the Theoretical Division at Los Alamos National Laboratory, have been chasing and observing these compact waves not on horseback, but by means of high-speed computation.

Although their findings are, so far, strictly mathematical, applications of Rosenau and Hyman’s compact waves may not be far off. For years, solitary waves have been considered as promising carriers of digital information on optical fibers, because they can, in principle, travel forever without losing their shape. Compact waves’ ability to travel close together without interfering with each other might offer even further advantages.

Mathematicians in the nineteenth century were slow to come to grips with Russell’s solitary wave, in part because the prevailing theory of wave motion was locked into a particular partial differential equation called the “wave equation,” which is still used to describe all kinds of undulatory phenomena, from water waves to sound waves to quantum-mechanical waves (the last, of course, being a twentieth-century innovation). According to the wave equation, Russell’s “heap of water” couldn’t sustain itself: It would immediately begin to break apart, as high-frequency components raced out in front, leaving lower-frequency components further and further behind—exactly what Russell didn’t see.

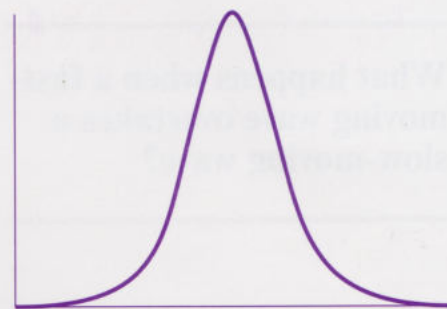
The wave equation also failed to explain another observation Russell made, this time when he re-created solitary waves in his laboratory by dropping weights into a long rectangular tank of water. The taller the wave, Russell found, the faster it moved. For explaining this behavior, the wave equation is no help at all: The wave equation is *linear*, and for phenomena described by linear equations, the height of things does not affect how they change in time.

By the end of the nineteenth century, however, an adequate theory for solitary waves had been developed, in the form of a modified wave equation known as the Korteweg–de Vries, or KdV, equation. Derived from basic equations of fluid dynamics, the KdV equation describes how waves propagate down a channel with



rectangular cross section. It differs from the ordinary wave equation in one critical respect: the KdV equation has a nonlinear term.

Rosenau and Hyman's equations go even further. The two theorists tinkered with the nonlinear term in the KdV equation; more important, they added a second nonlinearity, this time in a part of the equation known as the dispersion term (see box). The inspiration for making the dispersion term nonlinear came from Rosenau's studies of liquid drops, such as raindrops running down a window pane.



**Figure 1.** A traveling-wave solution of the KdV equation.

### A Tale of Three Equations

The wave equation, the KdV equation, and the compacton equation are all roughly similar in form, but the differences are critical. All three equations can be thought of as describing the up-and-down motion of a string of corks floating in a narrow channel of water, such as a long, thin trough. Mathematically, the trough is infinitely long and infinitely thin, so that each cork can be identified by a single variable, say  $x$ , which specifies its location along the length of the trough. The corks' up-and-down motion is described by a function of two variables:  $u(x, t)$  is the height of the cork at point  $x$  and time  $t$ .

The traditional, linear wave equation has the form  $u_t + u_x + u_{xxx} = 0$ . The first term,  $u_t$ , is the derivative of  $u$  with respect to  $t$ —that is, the speed at which a cork is going up or down. The middle term,  $u_x$ , is the derivative of  $u$  with respect to  $x$ ; it describes the slope of the wave at each point—that is, how much higher or lower each cork is than its neighbors at a particular moment. The final term,  $u_{xxx}$ , is the third derivative of  $u$  with respect to  $x$ . This is the “dispersion” term; because of it, traveling-wave solutions with different wavelengths propagate at different speeds. These solutions have the form  $u(x, t) = \sin((x - ct)/\ell)$  with  $c = (\ell^2 - 1)/\ell^2$ . The parameter  $\ell$  is the wavelength, while  $c$  is the speed at which the wave propagates (i.e., the speed at which a particular crest of the wave moves). Because the equation  $u_t + u_x + u_{xxx} = 0$  is linear, a general solution can be formed by adding these traveling-wave solutions together. But any such combination will produce a wave that changes shape over time, because the different components move at different speeds.

The KdV equation has the form  $u_t + (u^2)_x + u_{xxx} = 0$ . Changing the middle term—squaring the  $u$  before taking its first derivative—has a profound effect. Simple sine waves are no longer solutions; instead, the traveling-wave solutions have the form  $u(x, t) = (3c/2)\text{sech}^2((x - ct)\sqrt{c}/2)$ . The shape of the solution is a single, smooth hump (see Figure 1). What's more, both the height and the “width” of a wave are completely determined by the speed  $c$  at which it travels: Taller waves move faster than shorter ones, quite unlike waves governed by linear equations. Whatever shape a wave governed by the KdV equation has initially, it will eventually—and usually quite quickly—break up into a train of these basic shapes, with the tallest waves out front.

Rosenau and Hyman's compacton equations make the dispersion term in the KdV equation nonlinear as well. The simplest compacton equation has the form  $u_t + (u^2)_x + (u^2)_{xxx} = 0$ . (More generally, Rosenau and Hyman have studied equations of the form  $u_t + (u^m)_x + (u^n)_{xxx} = 0$ .) This time, the traveling-wave solutions have the form  $u(x, t) = (4c/3)\cos^2((x - ct)/4)$  for  $-2\pi < x - ct < 2\pi$  and  $u(x, t) = 0$  if  $|x - ct| \geq 2\pi$ . As with the KdV equation, the height of a compacton depends on its speed, but compactons have the same width, namely  $4\pi$ . Waves in any initial shape also decompose into a train of compactons. However, numerical evidence suggests that when compactons separate, as they do after a collision, they leave behind them an apparently infinite wake of tiny ripples. Rosenau and Hyman are still trying to fathom the nature of these ripples.

Most nonlinear equations cannot be solved exactly—that's one of the advantages linear equations hold. But some can. The KdV equation and the new nonlinear dispersion equations turn out to be among them. The basic traveling-wave solution of the KdV equation involves a special function known as a hyperbolic secant. (The



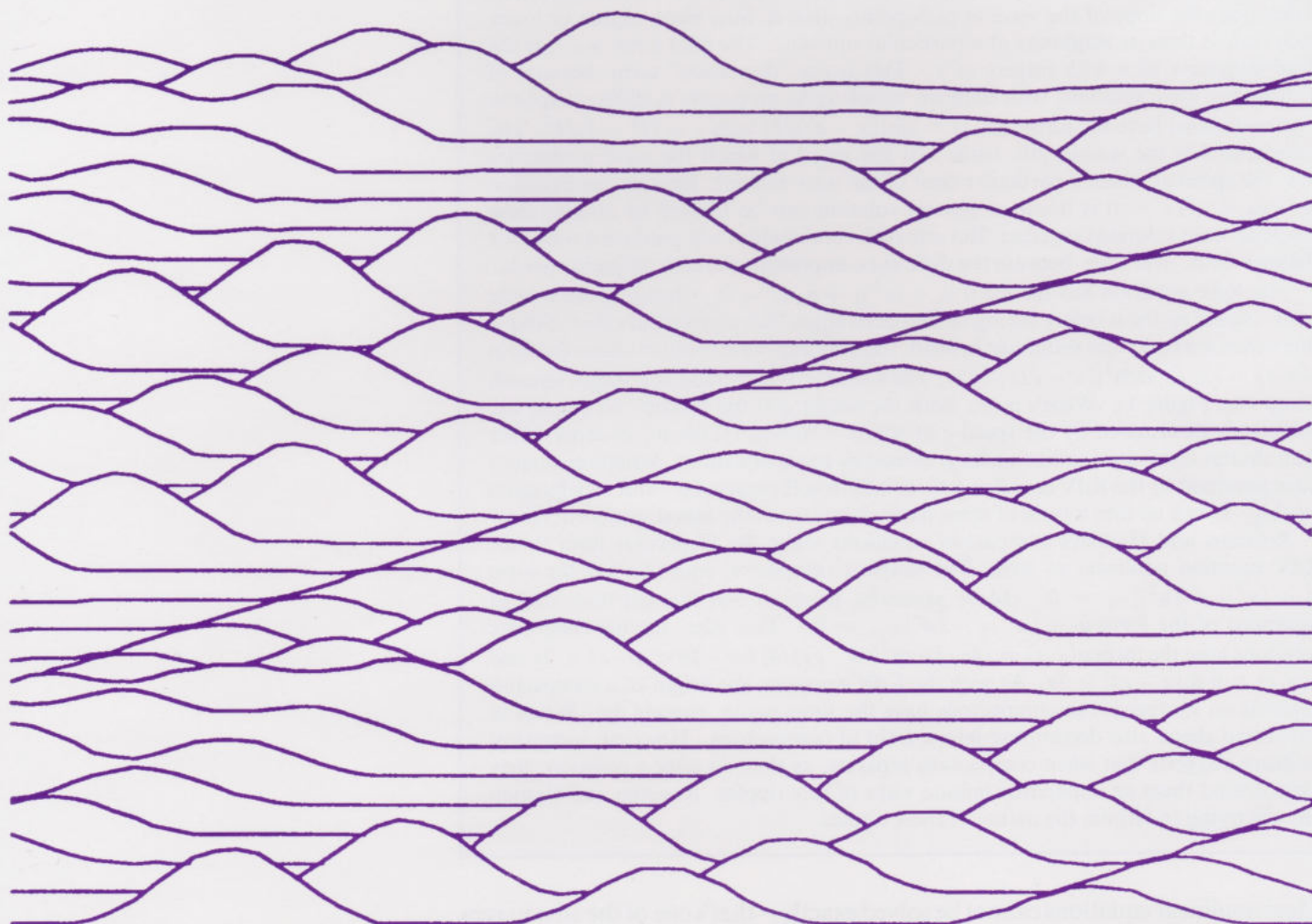
---

## What happens when a fast-moving wave overtakes a slow-moving wave?

---

hyperbolic functions, often seen in introductory calculus classes, are closely related to the trigonometric functions commonly studied in high school.) Rosenau and Hyman's equations, by contrast, have various traveling-wave solutions, ranging from parabolic arcs to cosine waves. But all of these waves, whether generated by the KdV equation or by Rosenau and Hyman's equations, share one particularly striking feature: The speed at which a solitary wave moves is proportional to its height—just as Russell had seen in his laboratory.

That feature raises an interesting question: What happens when a fast-moving wave overtakes a slow-moving wave? In the 1960s, Martin Kruskal at Princeton University and Norman Zabusky at Bell Telephone Laboratories in Whippany, New Jersey (both now at Rutgers University in New Brunswick, New Jersey), found a surprising answer. When a tall solitary wave overtakes a shorter one, the two do not merely merge. Nor do they break each other apart. Instead, after a brief but passionate encounter, the two waves separate, each with the same size and shape it had before. The only evidence they ever met is a “phase shift”: The taller wave is pushed slightly ahead of where it would otherwise have been, while the shorter wave is held slightly back. Because the solitary waves retain their separate identities, much as colliding particles do, Kruskal and Zabusky dubbed them “solitons.”



**Figure 2.** A space-time plot of three compactons colliding several times in a periodic domain. The compactons experience a phase shift with each collision and create a small ripple.



The property Kruskal and Zabusky discovered is not unique to the KdV equation. Many other nonlinear equations have the same property: In effect, their traveling-wave solutions (which are also called solitons) are impervious to any kind of disturbance. That makes solitons good candidates for carrying information. If, for example, light can be made to propagate along a fiber-optic cable in accordance with a KdV-type equation, then pulses of digital information can be sent as solitary waves, which travel long distances without distortion.

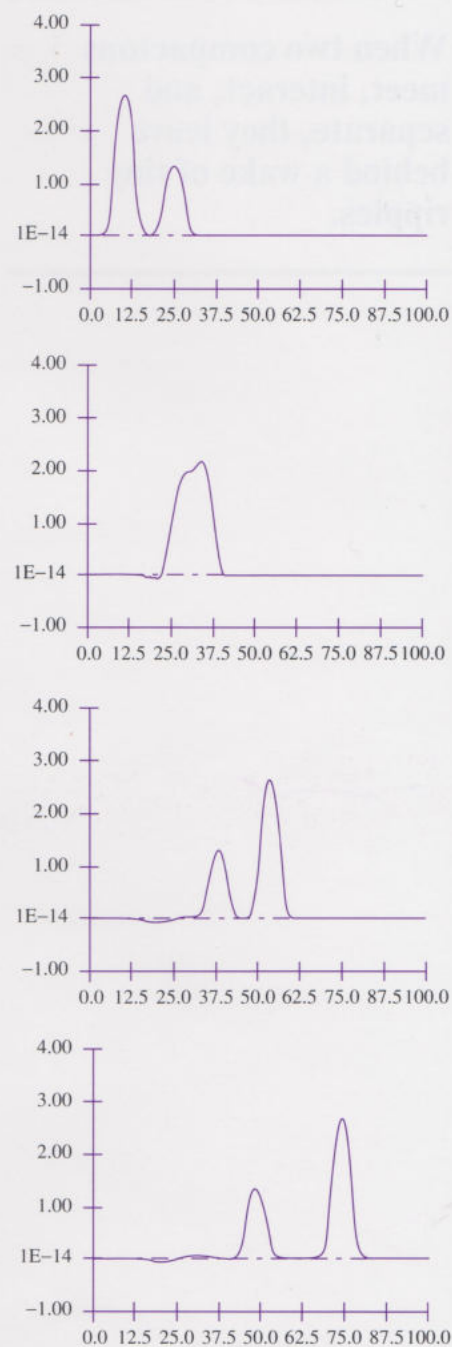
One drawback of KdV-type solitons, however, is that they aren't truly solitary. Each such "classical" soliton tapers off, both fore and aft, with an infinitely long "tail." As a result, two waves start interacting before their main parts meet (technically speaking, two waves are *always* interacting, but because the tails taper off exponentially, the interaction is weak until the waves are suitably close together). Thus to keep soliton-borne information from getting garbled, the carrier waves have to keep their distance.

That's where Rosenau and Hyman's compact solitons—or "compactons," as the two mathematicians call them—could have an advantage. These new waves are tailless: They vanish abruptly at the endpoints of a well-defined interval. As a result, two compactons cannot interfere with one another until they overlap. In theory, at least, a string of identical compactons could race along an information superhighway like so many manic tailgaters at rush hour—except that on this highway, everyone adheres strictly to the speed limit.

Whether compactons actually have a future on the fiber-optic highway remains to be seen. For now, Rosenau, Hyman, and their colleagues are interested mainly in the light compactons shed on the theory of solitons. One key insight regards the role of a condition known as integrability. The KdV and other classical soliton equations are all integrable. This means, roughly, that their solutions satisfy infinitely many "conservation laws," much as physical systems obey laws such as conservation of energy and momentum. Integrability helps explain solitons' extraordinary stability: The conservation laws constrain the waves so rigidly that they can hardly fall apart.

Compacton equations, however, are not integrable; they satisfy only a handful of conservation laws. Hyman didn't expect much to happen when they numerically smashed two compactons together, but Rosenau urged him to run the computer experiment. "We took one that's traveling fast and one that's traveling slow, and we banged them into each other," Hyman recalls. That the two waves emerged intact, just like ordinary, integrable solitons "was amazing," Hyman says. These unexpected results indicate that the remarkable stability of solitary waves lies deeper than mere integrability.

Still more surprising is a brand new feature, not seen in classical solitons: When two compactons meet, interact, and separate, they leave behind a wake of tiny ripples (see Figure 3). Rosenau and Hyman almost missed this in their first compacton calculations—they thought they saw only some numerical "noise" in the results, stemming from imprecisions in the computation. (All numerical computations are prone to round-off and other errors; one role of mathematical theory is to study such imprecision precisely). "It was only when we were getting ready to write up the results that we decided to do an extra-high-resolution calculation to get rid of this numerical noise," Hyman explains. "When it didn't go away, we started focusing in on it."



**Figure 3.** Ripples result when compactons collide. Here a tall wave overtakes a shorter one as they both move from left to right.



---

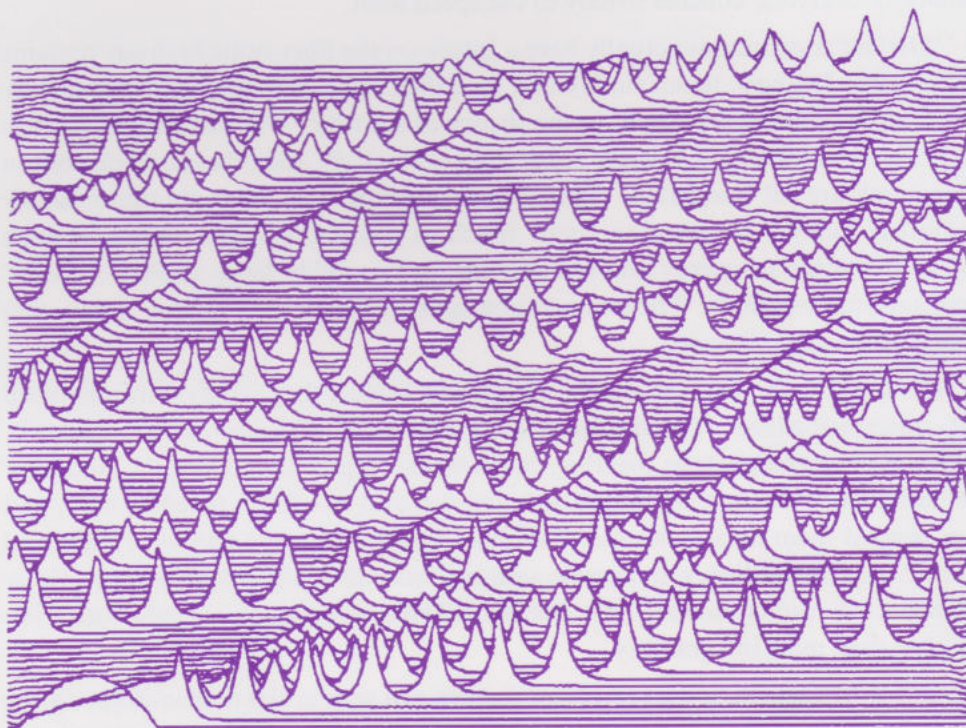
**When two compactons meet, interact, and separate, they leave behind a wake of tiny ripples.**

---

The ripples are a “real mystery,” Hyman says. They seem to continue indefinitely, with tinier and tinier ripples arising, a kind of flotsam caused, perhaps, by the compacton equations’ lack of integrability. But that’s just speculation: There’s no proof yet that the ripples don’t finally die out, just numerical evidence that smaller and smaller ripples continue to arise. “It’s just begging for a solution,” notes Hyman. Researchers may no longer chase chance observations on horseback, but plenty of “singular and beautiful” phenomena remain to be found.

### **A Peek at Peakons**

Hyman and colleagues Roberto Camassa and Darryl Holm at Los Alamos National Laboratory have also been looking at yet another new soliton-type equation. Like the compacton equations, this wave equation sports a nonlinear dispersion term, but the new equation also happens to be integrable. This time, the traveling-wave solutions have sharp peaks, hence the name “peakons” (see Figure 4). The researchers believe the peakon equation will provide additional insight into the role of nonlinear dispersion in the theory of solitons. Interestingly, the peakon equation was obtained by simplifying the equations of a global ocean circulation model—the same model that generated the color graphics for the cover of last year’s issue of *What’s Happening in the Mathematical Sciences*.



**Figure 4.** A space-time plot of peakons—a new kind of solitary wave—generated in this example from an initial parabolic hump (lower left).



# Mathematical Insights for Medical Imaging

**F**irst of all, do no harm." Along with his famous oath, the Greek physician Hippocrates left that instruction for his medical heirs. But in order to diagnose disease, modern physicians often find they must perform such invasive procedures as biopsy and angiography. Even X-rays are not without risk. It's a necessary evil: To treat disease, doctors need to see what's happening inside the body. Useful as it is, a stethoscope can't hear cancer cells growing.

Medical researchers are constantly looking for safer, more accurate ways to monitor patients' condition. A team of mathematicians and engineers at Rensselaer Polytechnic Institute (RPI) in Troy, New York, is doing its part to help. Mathematicians David Isaacson and Margaret Cheney, biomedical engineer Jonathan Newell, and their colleagues have developed a new, mathematically-based technology that produces real-time, continuous images of the heart, lungs, and other organs—all without cutting patients open or bombarding them with radiation. They hope their machine, which went into clinical testing at the Albany Medical

---

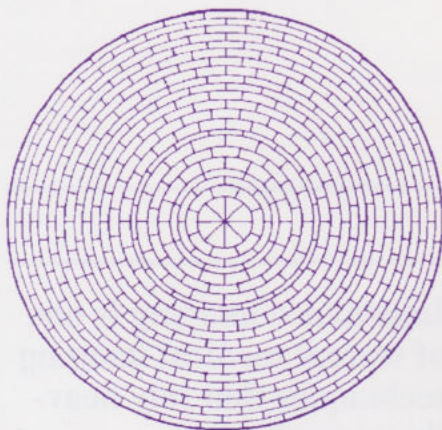
**Impedance imaging is one of several medical imaging techniques that rely heavily on mathematics.**

---



*Standing, left to right: Jonathan Newell, David Gisser, Gary Saulnier. Seated: Margaret Cheney and David Isaacson. In the left test tank, a piece of conducting pipe and a piece of insulating pipe are visible. These objects are placed in salt water that fills the tanks up to the top of the electrodes. The RPI group tests their impedance imaging system in these tanks. (Photo courtesy of the Rensselaer ACT group.)*





**Figure 1.** The 496-grid-cell mesh used to reconstruct images of electrical conductivity distributions. (Reprinted by permission of John Wiley & Sons, Inc., from "NOSER: An algorithm for solving the inverse conductivity problem," M. Cheney, D. Isaacson, J. C. Newell, S. Simske, and J. Goble, *International Journal of Imaging Systems and Technology*, vol. 2, p. 68, figure 1 (1990), © 1991 John Wiley & Sons, Inc.)

Center in 1993, will eventually offer physicians a powerful but safe diagnostic tool for such illnesses as heart disease, pulmonary edema, and breast cancer.

The new technology, known as Electrical Impedance Imaging, works by applying tiny electrical currents through electrodes placed on the skin, measuring the corresponding voltage response, and then deducing the distributions of electrical conductivity and permittivity inside the body. (Roughly speaking, conductivity measures how easily charge moves through a medium, while permittivity measures the capacity of a medium to store electrical energy.) Since different parts of the body have different electrical properties, the computed distributions provide an image of the body's tissues and fluids.

Take the lungs, for example. Air is a notoriously poor conductor of electricity. As a result, when healthy lungs fill with air, they show up in an impedance image as regions of low conductivity. By contrast, in a patient suffering from pulmonary edema—a complication often seen following injury, heart attack, or major surgery—the lungs are partially filled with fluid. Because the fluid has high conductivity, the edema appears as an abnormality in an impedance image.

Blood, too, has high conductivity, so impedance imaging also has potential for measuring the amount of blood being pumped by the heart. Measuring cardiac output "is very useful to physicians because it tells them how well the heart is working," explains Newell. "At present, the only reliable ways to measure cardiac output involve passing a catheter through a vein and through the heart, which is a dangerous and expensive procedure."

Impedance imaging is one of several medical imaging techniques that rely heavily on mathematics. The most common is Computed Axial Tomography, or CAT-scan. In essence, a CAT-scan combines X-rays taken from many different directions. Each X-ray measures the density of tissue along a particular line of sight. A computer algorithm based on a mathematical procedure called the Radon transform uses these measurements to reconstruct the actual spatial distribution of densities. Similarly, Magnetic Resonance Imaging, or MRI, constructs images by measuring the body's response to strong magnetic fields.

These techniques all involve solving what are known as "inverse" problems, so called because they ask, in effect, for the opposite of a direct calculation. If, for example, the conductivity distribution in an object is known, then the voltage response to a set of applied currents can be computed directly, much as an algebraic expression such as  $2x^2 + 7x - 5$  can be directly evaluated if the value of the variable  $x$  is known. On the other hand, the inverse problem—trying to reconstruct the conductivity distribution from a measured set of voltage responses—is like trying to find a value of  $x$  for which  $2x^2 + 7x - 5$  equals 2, only in a much more complicated mathematical setting.

For impedance imaging, the equations to be solved are derived from Maxwell's equations, a set of partial differential equations that describe all electromagnetic phenomena. Reconstructing an image of the body's interior from measurements on the surface is a considerable challenge, in part because the equations are nonlinear and in part because the reconstruction is highly sensitive to measurement errors. "Conductivity distributions that may be very different may produce data that are very close to each other," notes Isaacson. To cope with that problem, the RPI team has designed a high-precision electrical system for delivering current and measuring voltages, and coupled it with computer algorithms that optimize the system's performance.

The RPI group call their machine ACT III, for Adaptive Current Tomograph,



third generation. It combines delicate engineering with sophisticated mathematical analysis and high-speed computer algorithms to generate precise patterns of current and then reconstruct images from the measured responses. The currents, which are applied through electrodes like those used for electrocardiograms, are well below the level of human perception and considered harmless. That makes the system suitable even for continuous use, as a monitoring device. Whereas a CAT-scan, say, only takes “snapshots” of the body, impedance imaging makes movies, tracking physiological processes in addition to revealing anatomical structure.

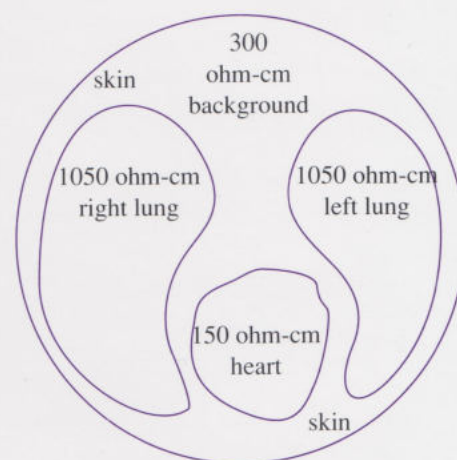
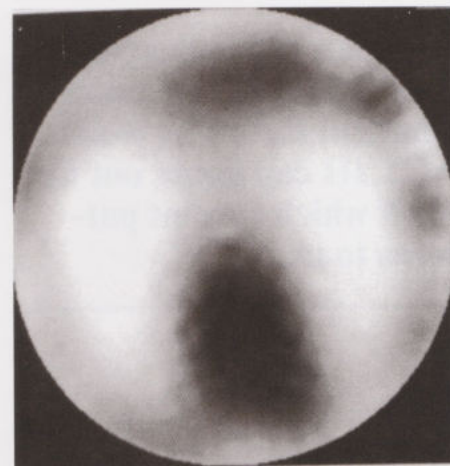
That’s not to say that impedance imaging will render CAT-scans obsolete. On the contrary, since they measure different properties of tissue, the two technologies complement each other. But impedance imaging offers some special advantages. For one thing, it’s relatively inexpensive, in part because of its compact electronics package. It also does not require a specialist to operate or interpret; ACT III or its likely successors could even be used by paramedics on ambulance calls.

Isaacson started studying the mathematics of impedance imaging in the early 1980s. He quickly saw that theory alone was not enough. “I had some ideas about things to do, but I needed some practical experience as to how accurately one can actually measure things, and I wanted to do some experiments,” he recalls. He went to Newell, who, while skeptical that impedance imaging could work in practice, helped design an experiment to find out. They decided to see whether Isaacson’s electro-mathematics could locate chunks of jello in a tray of saltwater.

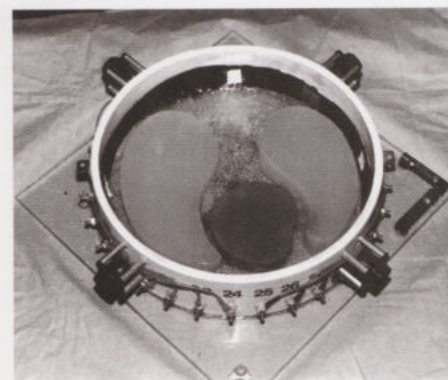
“We took a little pan from the local supermarket, filled it with gelatin and salt water, put electrodes around the outside of the pan, pumped some currents in and measured the voltages,” says Isaacson. Sure enough, an image appeared, which, though crude, showed roughly where the jello was. That was enough for Newell: “He got very excited about this,” Isaacson recalls.

Newell brought in David Gisser, an electrical engineer (now professor emeritus) at RPI, to design and build the electronics. The first system “was crude, but it worked,” Isaacson says. The current version incorporates many improvements in both hardware and software. In particular, Cheney notes, Gary Saulnier, an electrical engineer at RPI, and his student Peter Edic have made the system fast enough to work in real time. “A huge number of students have worked on the project at one time or another,” Cheney adds. “Ever since I’ve been associated with the project, there have been somewhere between 10 and 20 students involved at any one time—the number depends mainly on funding. They range from Ph.D. students to undergraduates. We’ve even had a couple of exceptional high school students.”

In experiments with electrodes surrounding a circular tray 30 centimeters in diameter (roughly the size and shape of a human chest), ACT III can reconstruct a reasonable image of a nickel-sized object in the center of the tray—the hardest spot to get a good picture. The RPI group is now also doing experiments with human subjects, including volunteer patients at the Albany Medical Center. To image a 2-dimensional “slice” through the heart and lungs, the researchers place a 32-electrode belt around a person’s chest. ACT III then sends a specially designed sequence of current patterns through the electrodes. Voltage measurements taken at the 32 electrodes are fed back to the machine, which uses an algorithm the group calls NOSER (Newton One-Step Error Reconstructor) to produce a circular, 496-grid-cell image (see Figures 1–3). The output is fed to a video monitor, on which the subject can watch—literally live—his or her own lungs filling and emptying



**Figure 2.** Impedance image (top) from a test with simulated heart and lungs having specified conductivities (bottom). (Photo courtesy of the Rensselaer ACT group.)



**Figure 3.** 30-cm test tank with simulated heart and lungs. (Photo courtesy of the Rensselaer ACT group.)



---

**Isaacson and colleagues have developed the mathematical theory by which ACT III can figure out for itself which current patterns to use.**

---

and blood pumping: Low-conductivity air appears in dark blue, high-conductivity blood—appropriately—in bright red.

Mathematically, each current pattern—say sending current in at just one electrode and taking it out at another—is a vector in 32-dimensional space. More precisely, each pattern is a vector in a 31-dimensional subspace defined by the requirement that the net current applied to the subject must be zero (otherwise the subject's hair would start standing on end). The measured voltage response at the electrodes is also a vector in 31-dimensional space. Roughly speaking, the conductivity distribution is to be found in the matrix that relates the current and voltage vectors. To find that matrix, it's necessary to apply 31 fundamentally different current patterns (in technical terms, the patterns must be "linearly independent").

One key question for the RPI group is which current patterns to use and how to design the electronics to get the best possible signal. In principle, any set that includes 31 linearly independent patterns will do. But that ignores the effect of errors, which can send the linear algebra rattling off into nonsensical solutions.

"It turns out that the best set of patterns to apply depends on what's inside the body," Isaacson explains. For imaging features near the body's surface, patterns that send current in at just one electrode and take it out at an adjacent electrode are optimal. The RPI group, however, uses patterns based on the trigonometric sine and cosine functions. These patterns are provably optimal for distinguishing features deep inside the body. Isaacson and colleagues have developed the mathematical theory by which ACT III can figure out for itself which current patterns to use.

The researchers are also exploring new reconstruction techniques. Cheney has led the way on one promising approach called layer stripping. Conceptually, layer stripping amounts to solving for the conductivity distribution layer by layer, like peeling an onion. The current and voltage measurements, which are made on the outside surface, are used directly to solve for the conductivity of the first layer. From this solution, a set of currents and voltages are computed for the *inside* surface of this layer. These "measurements" are then used to obtain the conductivity distribution of the next layer, and so on. "It's a simple idea," Cheney says. "The nice thing is, it applies to lots of problems."

The RPI researchers are not the only group working on impedance imaging, but Cheney credits Isaacson with having the clearest vision of what can be done. "One of the key things he is able to do is to ask the right questions," she says. "People working on inverse problems usually start by thinking about the reconstruction problem. Figuring out what data one needs in order to do reconstruction often suggests what measurements should be made. But Dave looked at the problem from the point of view of actually building a system, and asked the more fundamental question of how to make measurements containing the maximum amount of information."

The answers could wind up saving lives.



# Parlez-vous Wavelets?

**M**athematicians are like the French," the German poet Goethe once remarked. "They take whatever you tell them and translate it into their own language—and from then on it is something entirely different."

Goethe's observation is as true now as ever. But times may be changing. In the last ten years, mathematicians and researchers in diverse areas of science, engineering, and even art have discovered and begun to develop a theoretical language they can all understand. This new common language is sparking new collaborations. Many mathematicians are now crossing over into such applied areas as signal processing, medical imaging, and speech synthesis. At the same time, much deep but abstract-sounding mathematics is becoming accessible to researchers in fields from geophysics to electrical engineering.

The new language is wavelet theory. Those who speak it describe wavelets as powerful new tools for analyzing data. Wavelet theory serves as a kind of numerical zoom lens, able to focus tightly on interesting patches of data—but without losing sight of the mathematical forest while attending to the trees, twigs, buds, and grains of pollen.

"Never before in anything on which I've worked have I had contacts with people from so many different fields," says Ingrid Daubechies, a mathematician at AT&T Bell Laboratories and a leading authority on wavelet theory. Because there are so many aspects to the subject, "you have all these ideas brewing together—it's very fertile for everybody concerned," Daubechies adds. "It's a very nice laboratory for showing that applications can have interest for pure mathematics, and vice versa."

Mathematically, wavelets are an offshoot of the theory of Fourier analysis. Introduced by the French mathematician Joseph Fourier in his essay *Théorie analytique de la chaleur* (analytic theory of heat), published in 1822, Fourier analysis seeks—with great success—to understand complicated phenomena by breaking them into mathematically simple components. The fundamental idea is to take a function and express it as a sum of trigonometric sine and cosine waves of various frequencies and amplitudes. The familiar and well-understood trigonometric functions are easy to analyze. By combining information about a function's sine and cosine components, properties of the function itself are easily deduced—at least in principle.

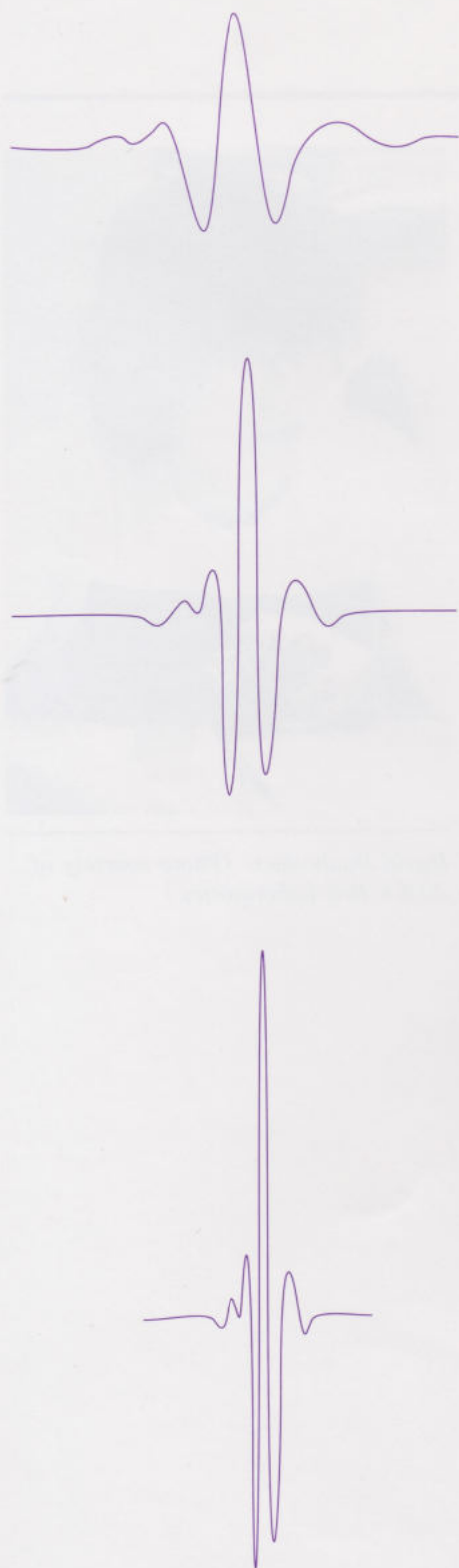
Fourier analysis is among mathematics' most widely used theories. It is especially suited to analyzing periodic phenomena, periodicity being the most prominent property of sines and cosines. But even so, the theory has its limitations and its pitfalls. The main problem is that finding detailed information about a function requires looking at a huge number of its infinitely many Fourier components. For example, a transient "blip," obvious in a graph, is impossible to recognize from its effect on a single component. The reason, in essence, is that each sine and cosine wave undulates infinitely in both directions; thus a single wave can't help locate anything. Indeed, the sharper the blip, the more Fourier components are needed to describe it.

Wavelet theory takes a different approach. Instead of working with the infinitely undulating sine and cosine waves, wavelet analysis relies on translations and dilations of a suitably chosen "mother wavelet" that is concentrated in a finite interval. Almost any function can serve as the mother wavelet; this makes wavelet theory



Ingrid Daubechies. (Photo courtesy of AT&T Bell Laboratories.)





**Figure 1.** A “mother” wavelet (top) and two “daughters.” (Figure courtesy of Ingrid Daubechies.)

more flexible than traditional Fourier analysis. “Daughter” wavelets are formed by translating, or shifting, the mother wavelet by unit steps and by contracting or expanding it by powers of two (see Figure 1). One then expresses other functions as combinations of wavelets, just as Fourier analysis represents functions by combining sines and cosines.

The fact that the mother wavelet is concentrated in a finite interval gives wavelet theory its zoom-in capability: An interesting blip in a function can be analyzed by looking only at those wavelets that overlap with it; finer details are resolved by looking at increasingly contracted copies of the mother wavelet in the vicinity of the blip.

Many of the ideas underlying wavelet theory have been around for decades, but the subject itself got off the ground only recently. The story starts in the early 1980s in France, when wavelets were introduced by geophysicist Jean Morlet and mathematical physicist Alexander Grossmann. In 1985, mathematician Yves Meyer constructed a family of wavelets with two highly desirable mathematical properties, called smoothness and orthogonality. (Interestingly, J. O. Stromberg at the University of Tromsø in Norway had constructed such a family several years earlier, but the connection with the nascent theory of wavelets was not realized until after Meyer’s work.)

The following year, Meyer and Stephane Mallat gave the subject a solid foundation with a theory of “multiresolution analysis.” Then in 1987, Daubechies constructed a family of wavelets that, in addition to being smooth and orthogonal, were identically zero outside a finite interval. Daubechies’s construction opened up the field. “Compactly supported” wavelets are now easy to come by, and are among the most commonly used in applications.

And applications are abundant. Wavelets are being tested for use in everything from digital image enhancement—making blurry pictures sharp—to new methods in numerical analysis (itself widely used in scientific computing). “They’re a very versatile tool,” says Daubechies. Not all the applications will pan out, but many will, and some already have. “There are some very nice success stories,” Daubechies adds.

One such story may have far-reaching effects, especially for the next generation of criminals. The Federal Bureau of Investigation has adopted a wavelet-based standard for computerizing its fingerprint files. The FBI has around 200 million fingerprint cards on file, according to Peter Higgins, deputy assistant director of the Bureau’s Criminal Justice Information Services division, and 30,000 to 40,000 identification requests pour in every day. At present, the FBI’s fingerprint files consume about an acre of office space. The goal, says Higgins, is to digitize the files, store them electronically, and “put [them] in something that would fit in a 20 × 20-foot room.”

It sounds easy; after all, entire encyclopedias now fit on a compact disk with room to spare. But that’s words. Images are something else. At a resolution of 500 pixels per inch, a standard fingerprint card contains nearly 10 megabytes of data. Transmitting that much information over a modem—something the police would like to be able to do—takes hours at today’s transmission rates. For a dozen cards, it’s quicker to use Federal Express.

What’s needed is some way to compress the data on a fingerprint card without distorting the picture. That’s where wavelets come in. By treating the fingerprint image as a two-dimensional function, it’s possible to represent it with a combination of wavelets. With a suitably chosen family of wavelets, only a relative handful

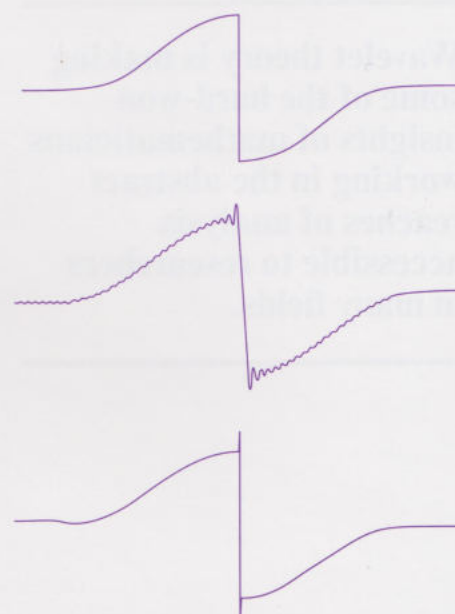


are needed to represent a fingerprint, and the contribution of each wavelet can be rounded off, or “quantized,” which reduces the amount of data that needs to be stored or transmitted.

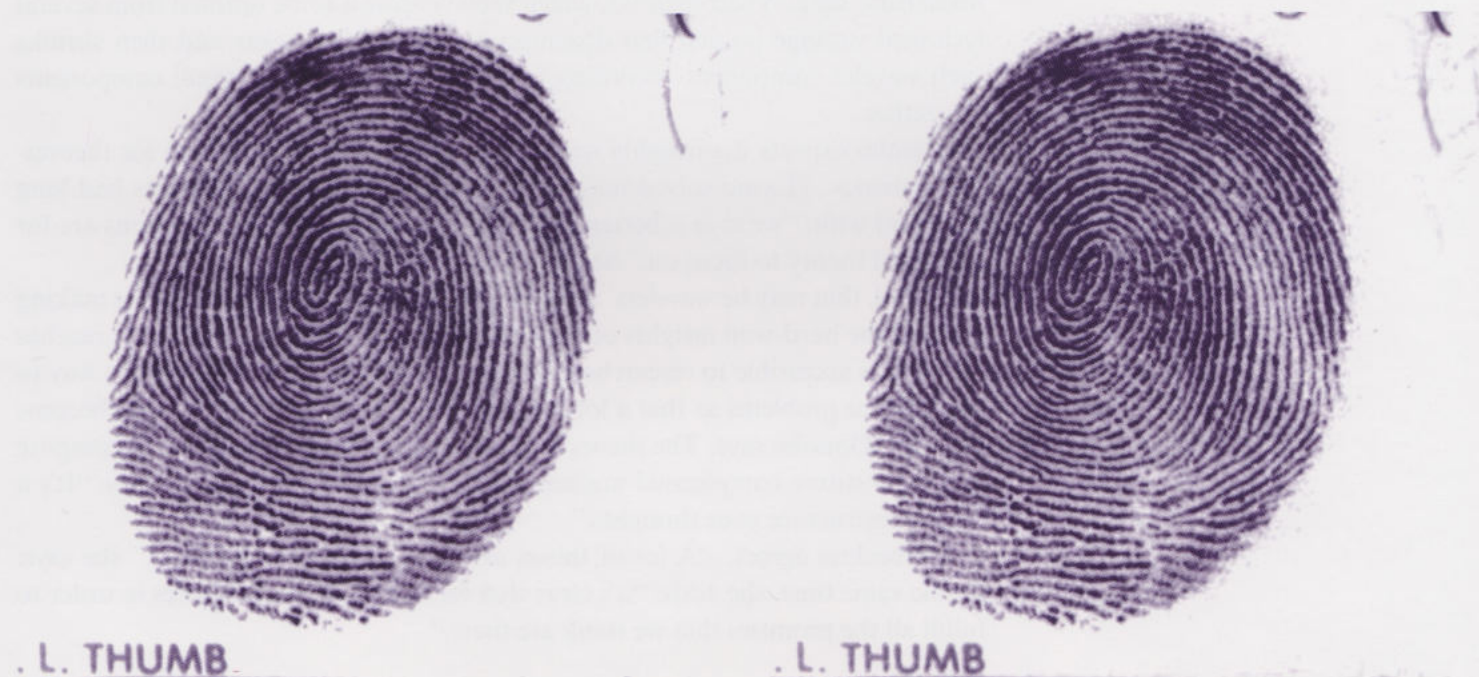
The wavelet standard for fingerprints was developed by Tom Hopper at the FBI and Jonathan Bradley and Chris Brislawn at Los Alamos National Laboratory. The standard allows many kinds of wavelets to be used—in effect, each electronic fingerprint “card” will include formulas for its particular wavelets, as well as the wavelet representation of the fingerprint itself. So far, one family of wavelets has been approved for use. It compresses fingerprint data by a factor of approximately 20 to 1—reducing 10 megabytes to a much more manageable 500 kilobytes—yet gives images that pass the FBI’s automated recognition tests. Indeed, the reconstructed fingerprints look almost exactly like the originals (see Figure 3).

Bradley and Brislawn have also applied wavelet techniques to another kind of data compression: managing the numerical geysers that gush out of supercomputers when running such things as global climate models. “High-performance computers are reaching the point where their ability to churn out data is surpassing our capacity for storing and analyzing it,” says Brislawn. In the approach he and Bradley have developed, the computer decomposes the solution (for example, a color-coded map of global ocean temperatures) into wavelets; the user can then control the output by specifying how much detail—that is, how many of the wavelet components—he or she wants to see. One challenge is to figure out how much you can compress the output without sacrificing quantitative capabilities of a model, such as long-term statistical predictions of climatic conditions, Brislawn notes. “This looks like a tough question that we won’t be able to answer until we get a better idea of what the models are capable of predicting.”

Other researchers are studying the use of wavelets not as post-processing tools, as Bradley and Brislawn are doing, but directly in scientific computation itself. Gregory Beylkin at the University of Colorado has been studying applications



**Figure 2.** A Fourier (middle) and wavelet (bottom) reconstruction of a function (top) with a sharp discontinuity. The Fourier reconstruction uses 65 nonzero coefficients, the wavelet reconstruction only 18. (In both cases, the discontinuity causes an overshoot, known as a Gibbs phenomenon, but it is much more localized in the wavelet reconstruction.)



**Figure 3.** (Left) Original  $768 \times 768$  8-bit Gray-scale fingerprint image. (Right) Fingerprint image compressed 26.0:1. (Photos courtesy of Chris Brislawn, Los Alamos National Laboratory.)



---

**Wavelet theory is making some of the hard-won insights of mathematicians working in the abstract reaches of analysis accessible to researchers in many fields.**

---

of wavelets in numerical analysis. Many problems—such as solving a system of partial differential equations that describes the flow of oil underground—boil down to working with huge matrices, or square arrays of numbers. Such matrices are easier to work with if many of their entries are zero. Beylkin has shown that wavelet analysis can reduce a wide class of matrices to the desired form.

Wavelets are especially suited to analyzing sound. Indeed, there's a strong resemblance between wavelets and musical notes. The mother wavelet can be likened to a particular note—say a quarter note at middle C—played at a particular time. Its translates represent the same quarter note at middle C played at other times, while its contractions and expansions are eighth- and half-note C's, played at higher and lower octaves. Ronald Coifman at Yale University and Victor Wickerhauser at Washington University in St. Louis have developed a technique they call adapted waveform analysis, in which a catalog of waveforms is automatically searched for the wavelets best suited to a particular problem. Among the applications is removing noise from recorded sound.

Coifman and his colleagues recently cleaned up an old piano recording of Johannes Brahms playing one of his own Hungarian Dances. Over the years, the recording had acquired several layers of noise. Brahms's live performance was recorded in 1889 on a wax cylinder, which later partially melted. The damaged cylinder was re-recorded on a 78 rpm disk; the version Coifman began with had been recorded from a radio broadcast of the 78. By then the music, competing with pops, hiss, and static, was all but inaudible. Wavelet techniques made it possible to remove enough noise to hear Brahms playing.

Wavelets are also helping researchers clean house in theoretical statistics. "As soon as we were exposed to wavelets, we made the equivalent of about ten years' progress in months," says David Donoho, a Stanford University statistician who has led the way in applying the new theory. Donoho and his colleague Iain Johnstone have developed a "wavelet shrinkage" method for removing numerical noise from data. Their method, which they've shown to be optimal from several technical vantage points, first decomposes data into wavelets and then shrinks each wavelet component according to a rule that eliminates small components altogether.

Donoho expects the insights wavelets supply to set a new agenda for theoretical statistics. Having solved many of the technical problems theorists had long struggled with, "we're in a better position to say what the right questions are for statistical theory to focus on," he says.

Indeed, that may be wavelets' most important legacy. Wavelet theory is making some of the hard-won insights of mathematicians working in the abstract reaches of analysis accessible to researchers in many fields. "Wavelets teach you a way to think about problems so that a lot of ideas in abstract harmonic analysis become natural," Donoho says. The theory of wavelets does more than simply decompose and reconstitute complicated mathematical functions. In Donoho's view, "It's a tool to restructure your thoughts."

Daubechies agrees. "A lot of things are starting to come together," she says. At the same time, she adds, "it's clear that we still need new advances in order to fulfill all the promises that we think are there."



# Random Algorithms Leave Little to Chance

It's a common experience: You're walking down an office corridor or a city sidewalk when, without warning, you find yourself face to face with someone in an equal hurry going the other way. You both stop before you collide, and you both step aside—to *your* right. You both smile awkwardly and both step aside again—to your *left*. You both smile again. This time you wait for the other to make a move. You *both* wait for the other. Finally one of you breaks the pattern, and the impasse ends. You both laugh, say "thanks for the dance," and walk away wondering how long you could have both been stuck there.

That scenario generally plays out to comic effect in everyday life. Curiously, something similar occurs in computers—but with effects that are less amusing. When a single-minded program meets the wrong input, the result can be a devastating slowdown. And according to Murphy's Law ("If anything can go wrong, it will"), if there's a data set on which a program runs slowly, then that's the data set the program will be asked to process.

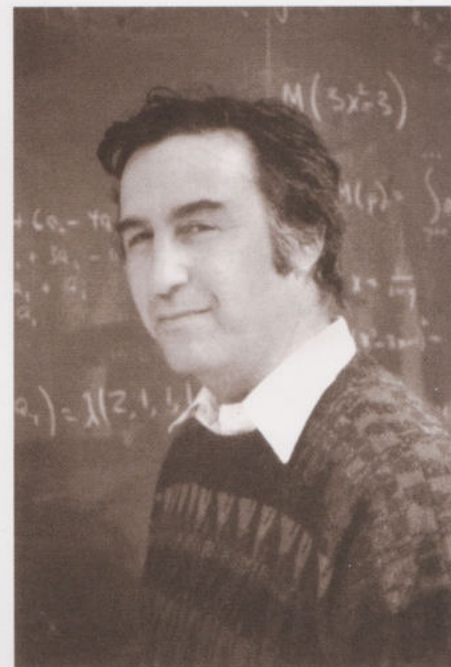
There may be a way out, though. Mathematicians and computer scientists are studying a new approach to programming that avoids the computational gridlock associated with many problems. This new approach relies on a humble but time-honored technique: flipping coins.

The technical term is "randomization," but it boils down to heads and tails. The idea is to insert occasional random decisions into a computation to avoid getting caught up in some unexpected conspiracy between program and data. While random algorithms are susceptible to runs of bad luck, such runs can be made exceedingly improbable. Moreover, that kind of bad luck is independent of the data.

"When you put coin flips into your algorithm, then it doesn't matter how your data is structured," explains Joel Spencer of the Courant Institute of Mathematical Sciences at New York University. "There's no particular kind of data that's bad."

Here's how the idea works in the case of the sidewalk tango. Suppose that you, the program, have a strictly deterministic pattern of responses to the other pedestrian (the data, which in this case is another program). If, say, you always cycle through the responses Left, Right, Wait, then you'll be OK if the data has some other pattern. But if the data happens to be structured the wrong way, then you're stuck forever. On the other hand, if you randomly choose among the possible actions each time, then no matter how the data is structured, it is highly unlikely you'll be blocked for long. Even if the data "wants" to block you, it can't—unless it's somehow clairvoyant, in which case you've got bigger problems. (It's also possible you've stumbled across a mirror.)

Computers, of course, rarely go walking down the sidewalk. A more realistic setting where randomness helps is the task of sorting. Computers are often fed long lists (of names, say, or addresses) to be put into alphabetical or numerical order: it's the kind of "mindless" activity computers excel at. And that's the problem: A machine will gladly spend all day—or all decade—sorting census data, and it will do just that if you don't worry about the efficiency with which it works.



Joel Spencer. (Photo courtesy of Barry Cipra.)



The efficiency of a sorting algorithm (what computer scientists call its “computational complexity”) is measured by the number of pairwise comparisons it makes—that is, how often the algorithm compares two objects to see which one comes first. If an algorithm is unlucky (or stupid) it can wind up comparing every pair of items. That’s not so bad if you’re just trying to put a bridge hand in order. But it’s a grim prospect if you’ve just spilled a thousand alphabetized index cards onto the floor—you could wind up making nearly half a million comparisons. And when the number of items, say on a political party’s mailing list, climbs into the millions or tens of millions, the potential worst-case number of comparisons begins to make the national debt look like a pittance.

One popular sorting algorithm is called QuickSort. The basic idea is to choose one item on the list, such as the item currently on top, and then compare everything else with it, forming two piles: those “above” and those “below.” The key then is to repeat this procedure with the “above” and “below” piles *separately*—there is no need ever to compare items from different piles. This process, which theorists call “recursive,” is guaranteed to work.

On most lists, QuickSort works quite well. More precisely, when averaged over all possible arrangements of a list, the number of comparisons the algorithm makes is proportional to the number of items in the list multiplied by the logarithm of that number. But there are times when QuickSort doesn’t work well at all. Ironically, if the list is *already* sorted, then QuickSort does the worst possible thing: It compares every pair of items.

“It’s not enough that an algorithm does well on average if there’s a patch of problems on which it does very badly, and that patch of problems happens to come up in the real world,” says Spencer. “That’s exactly the case with QuickSort, because in the real world you *do* sort things that are already sorted.”



**Figure 1.** QuickSort arranges a bridge hand. In this example, the bridge hand as dealt (top) is rearranged into proper ascending order (bottom) with 30 comparisons. In the worst case, QuickSort makes 78 comparisons to get a hand in order.



A randomized version of QuickSort solves the problem of “bad” data: Instead of starting with the first item on the list, or any other predetermined item, pick an item *at random*. Most of the time, the two piles will be of roughly the same size. By choosing a random item at each stage, the algorithm will—unless you are exceedingly unlucky—make close to an average-case number of comparisons in total.

QuickSort, whether randomized or not, always produces a correct answer—that is, a properly sorted list. What you’re gambling on is not the answer, but how long it takes the algorithm to find it. In other applications, an algorithm’s run time is guaranteed, but the answer it produces is only approximate, but with a high probability of being very close. One such problem concerns computing (or estimating) the “volume” of an  $n$ -dimensional shape. This is not just an arcane mathematical pursuit; the size of higher-dimensional geometric shapes is a central concern in many problems in theoretical physics, chemistry, statistics, and elsewhere.

Theoretical computer scientists have proved that estimating the volume of an  $n$ -dimensional shape to a specified accuracy is computationally intractable, if the algorithm used is deterministic. “Intractable” means that the amount of computation increases exponentially with the dimension  $n$ , leading to a kind of computational inflation that makes all but the smallest problems too expensive to solve. In general, theorists consider a problem tractable if there is a “polynomial time” algorithm for solving it—that is, an algorithm whose computational demands increase no faster than some power of the size of the problem (in this case, the dimension  $n$ ). For estimating volume, there is no such algorithm.

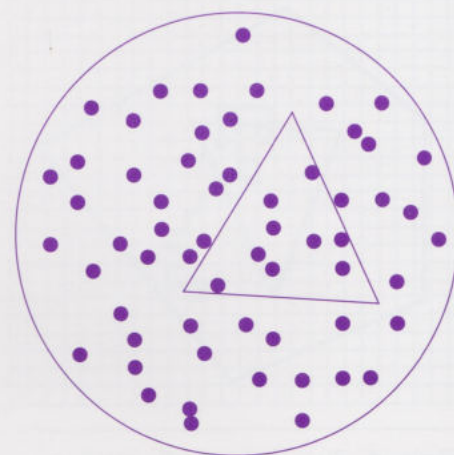
No such *deterministic* algorithm, that is.

In 1989, Martin Dyer at the University of Leeds in England and Alan Frieze and Ravi Kannan at Carnegie Mellon University found a random algorithm for estimating volume that broke through the problem’s exponential barrier. Their algorithm, which computes the volumes of convex bodies, is based on a method for quickly “getting lost” inside an  $n$ -dimensional shape.

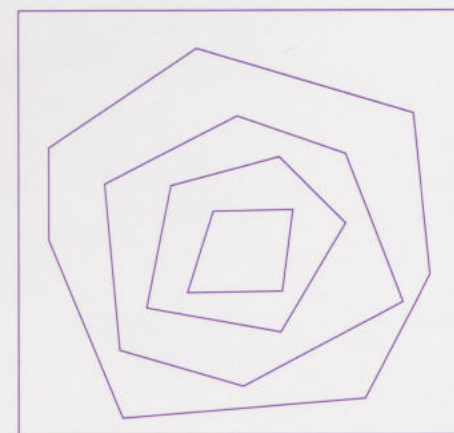
The starting point for Dyer, Frieze, and Kannan’s algorithm resembles a poorly played game of darts: If you throw darts without aiming, the fraction that hit a particular region of a dartboard is approximately equal to that region’s fraction of the dartboard’s area (see Figure 2). Curiously enough, an accurate estimate for the region’s area requires an *inaccurate* aim.

In  $n$  dimensions, the traditional “dartboard” is an  $n$ -dimensional “cube,” and “darts” are thrown by picking a random number for each of the  $n$  coordinates of a point in the cube. But that alone doesn’t solve the problem. Estimating volume this way requires a number of darts that grows exponentially with  $n$ . The reason is somewhat counter-intuitive: An object can fit snugly into the  $n$ -dimensional cube but still occupy just an exponentially tiny portion of the cube’s volume. For example, the  $n$ -dimensional “sphere” of diameter 1 touching all sides of a unit cube has volume less than  $1/2^n$  if  $n \geq 12$  (see box next page). Therefore, to have any reasonable chance of estimating the volume of, say, a 100-dimensional sphere, you’d have to throw more darts than there are elementary particles in the universe.

The three theorists dodge that problem by placing the “target” region inside a nested set of dartboards. The dartboards—really, just convex shapes in  $n$ -dimensional space—are crafted so the target occupies a substantial fraction of the smallest, which occupies a substantial fraction of the next smallest, and so on, until the largest dartboard occupies a good bit of the cube (see Figure 3). By randomly

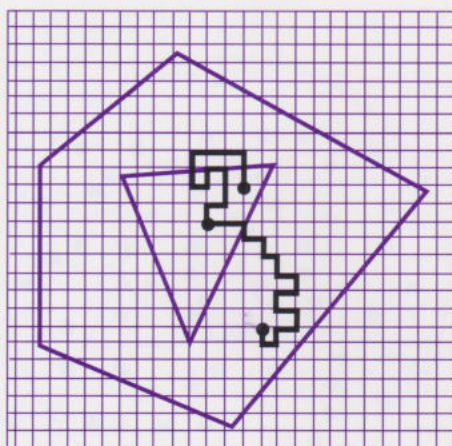


**Figure 2.** Nine out of 60 randomly thrown “darts” land inside the triangle, providing an estimate of the triangle’s size relative to the circle



**Figure 3.** A nested set of “dartboards” can be used to estimate the size of shapes.





**Figure 4.** Two random walks inside a pentagon. Both start at the same point in the triangle, but only one terminates inside the triangle.

throwing darts at the smallest dartboard, you can estimate the target's volume as a fraction of that board. Likewise, each dartboard's volume as a fraction of the next larger board can be estimated by randomly throwing darts at the larger board. The final result—an estimate for the volume of the target as a fraction of the cube—is obtained by multiplying all these fractions together.

### The Incredible Shrinking $n$ -sphere

The area and volume formulas  $\pi r^2$  and  $(4/3)\pi r^3$  are familiar to anyone who has studied circles and spheres. Less familiar, perhaps, is the formula

$$\frac{\pi^{n/2} r^n}{\Gamma((n/2) + 1)},$$

which gives the “volume” of an  $n$ -dimensional “sphere” of radius  $r$ .

The denominator,  $\Gamma((n/2) + 1)$ , takes some explaining. The gamma function, as it's called, is a much-studied special function. It plays important roles throughout mathematics, from geometry to number theory. The gamma function generalizes the factorial function  $n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1$ —itself a central character in combinatorics and probability theory. For computational purposes, the main property of the gamma function is the “recursion relation”  $\Gamma(x+1) = x\Gamma(x)$ . Thus, for example,

$$\Gamma(4) = 3 \cdot \Gamma(3) = 3 \cdot 2 \cdot \Gamma(2) = 3 \cdot 2 \cdot 1 \cdot \Gamma(1).$$

Likewise,

$$\Gamma\left(\frac{7}{2}\right) = \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \cdot \Gamma\left(\frac{1}{2}\right).$$

To round out the application to the  $n$ -sphere, it's enough to know that  $\Gamma(1) = 1$  and  $\Gamma(1/2) = \pi^{1/2}$ . Thus, for example, the volumes of the 4-, 5-, and 6-spheres of diameter 1 (radius  $1/2$ ) are  $\pi^2/32$ ,  $\pi^2/60$ , and  $\pi^3/384$ , respectively. In general, the volume of the  $n$ -sphere gets smaller as  $n$  gets larger: The numerator  $(\pi/4)^{n/2}$  decreases exponentially, while the denominator  $\Gamma((n/2) + 1)$  increases. As a result, even though the  $n$ -sphere fits snugly inside a cube, the fraction of the cube's volume that it occupies is exponentially small—making it a tiny target if you try playing “darts” on the cube (see main story).

But that strategy only trades one problem for another that seems equally difficult: picking random points inside an arbitrarily shaped region.

Again, if you're not concerned with doing things quickly, there's no problem: All you have to do is use random numbers to generate the coordinates of points in  $n$ -dimensional space but discard any points that happen to fall outside the desired region. To keep the computation tractable, however, Dyer, Frieze, and Kannan had to find another strategy—and then prove that it works.

The approach they found involves one of the staples of probability theory: random walks. The new strategy doesn't produce truly random points, but it comes close enough for many purposes, including the  $n$ -dimensional volume-estimation problem. The random walks take place on an  $n$ -dimensional grid (see Figure 4). Starting from a point that's known to be in the region of interest, the random walker picks one of the  $2n$  coordinate directions at random (in three dimensions, for example, she might roll a die to decide whether to go back, forth, left, right, up, or down) and then moves one step in that direction—provided doing so doesn't take her outside the region.

Researchers have known for a long time that a random walker eventually “gets



lost” in the sense that after a certain number of steps she has nearly equal probability of being found at any given grid point. The open question was how long it takes to get lost—does the number of steps grow exponentially or polynomially with the dimension  $n$ ?

“We showed we can get lost in polynomial time,” explains Kannan. In other words, even though the number of grid points grows exponentially with the dimension  $n$ , the number of random steps it takes to get anywhere on the grid with nearly equal probability grows no faster than some power of  $n$ . To prove it, “we needed a fair bit of mathematics,” including “various results from differential geometry that had just been proved in the 1980s,” Kannan says.

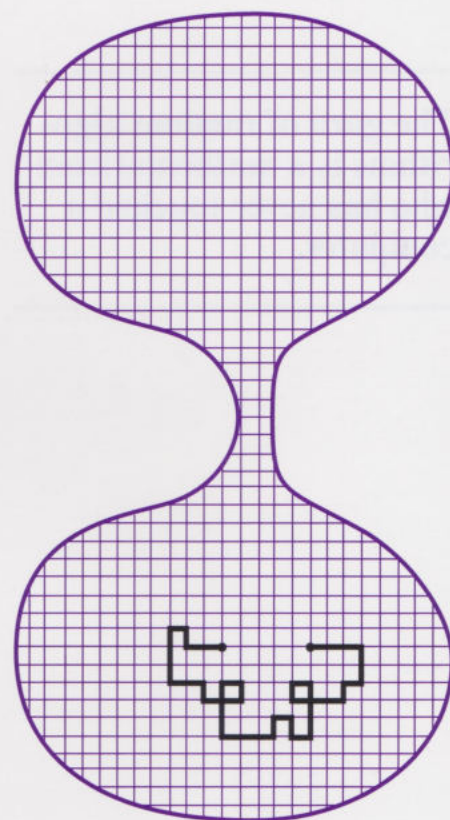
Dyer, Frieze, and Kannan’s result applies only to convex regions—and even then, some technical restrictions apply. (It’s not hard to see why the method doesn’t work in general: If your region is hourglass-shaped, as in Figure 5, then a random walk starting in one compartment may require exponentially many steps to “discover” the other compartment.) The three theorists’ original proof showed that the amount of computation required for accurate volume estimates in  $n$  dimensions is bounded by  $n^{27}$ —a marked improvement over the exponential bounds of deterministic algorithms, but still far from practical. (By contrast, the worst-case behavior of deterministic QuickSort is bounded by  $n^2$ .) Subsequent work by a number of researchers, though, has lowered the bound. Most recently, Kannan, László Lovász of Yale University and the Eötvös Loránd University in Budapest, and Miklós Simonovitz of the Hungarian Academy of Science have introduced techniques that lower the bound to  $n^5$ —and, if a certain conjecture is true, down to  $n^4$ . The algorithm, which could have a multitude of applications, is now “verging on the practical,” Kannan says.

It may be some time before random algorithms become commonplace in computer applications. “There’s a real preference among many people in the real world for deterministic algorithms,” Spencer acknowledges. But the theory is burgeoning, and the potential is very real. Says Spencer: “There’s something that’s not coincidental in the effectiveness of randomized algorithms. They’re not just a quirk. I think they’re really important for computer science.”

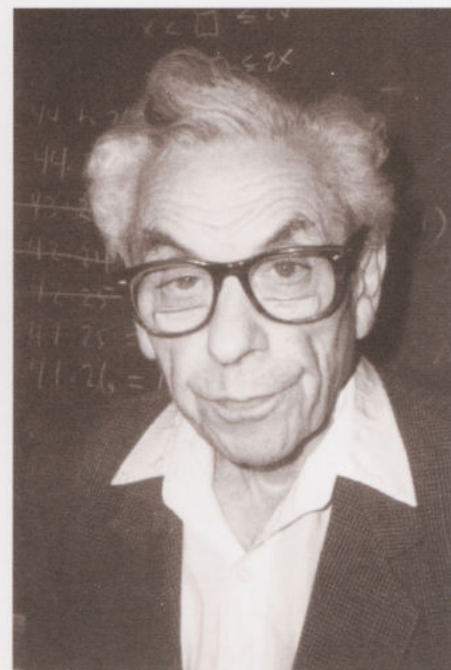
### The Guru of Random Algorithms

The guru of random algorithms is a mathematician who has never touched a computer: Paul Erdős, one of the best-known and most colorful mathematicians of the twentieth century. Erdős, who turned 80 in 1993, is a frequent visitor to research centers around the world. He has written hundreds of papers and co-authored many hundreds more. Joel Spencer credits him with invigorating the theory of combinatorics and creating what Spencer calls the probabilistic method, which, while purely mathematical, is what makes random algorithms tick.

One of Erdős’s specialties is proving the existence of combinatorial structures without actually constructing them. The probabilistic method, for example, does this by showing that, under the right circumstances, an object picked at random from a certain class of combinatorial objects will have the desired structure with probability greater than zero—and that can happen only if objects with the desired structure exist.



**Figure 5.** A random walk in a nonconvex region may take a long time to “explore” the whole region.



Paul Erdős. (Photo courtesy of Barry Cipra.)

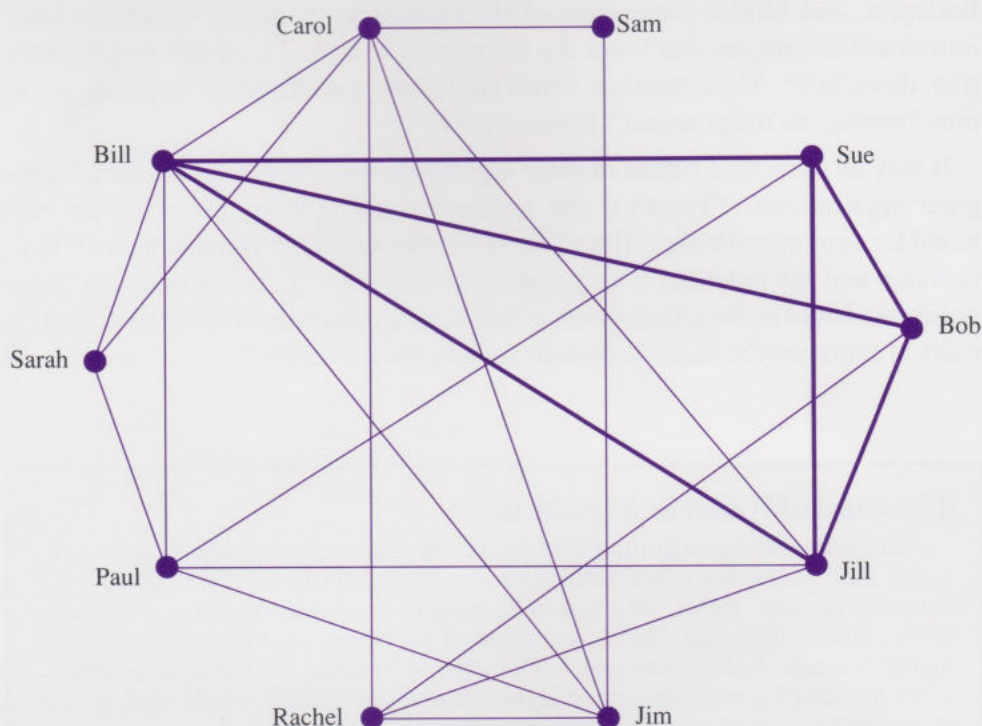


**The guru of random algorithms is a mathematician who has never touched a computer.**

Spencer's favorite example is Erdős's very first: a result in graph theory dating back to 1947. The problem is easiest to state in terms of social engineering: Is it possible to throw a party for, say,  $n$  guests, at which there are no large groups either of mutual friends or of mutual strangers? (To be precise, by "large" we mean twice the logarithm base 2 of  $n$ .) Erdős's answer: Yes. Just start with a roomful of mutual strangers, bring each pair together and either introduce them or not, depending on the toss of a coin. By an ingenious proof, Erdős showed that the probability of getting a party with the desired mix is not just greater than zero, it's extremely close to certainty.

That might seem to suggest that Erdős's random introductions could be replaced by some deterministic rule. But so far, no one has found one that works. (The problem, of course, is to find a rule that works for *all* values of  $n$ .) "No one has even come close to this result by a constructive [algorithm]—and it's been 46 years," says Spencer. The reason, he speculates, is that "when you start to construct things, you're putting structure into them, and this problem seems to demand a *lack* of structure." But who knows? One of Erdős's many protégés might still find a construction that solves the problem.

That's happened with other problems. Indeed, "derandomization" is a hot topic in the theory of random algorithms. Spencer notes. It isn't always possible, and it often makes the inner workings of an algorithm harder to understand, but derandomization offers theoretical insights of its own. At the very least, it gives theorists something to wager over.



**Figure 6.** Ten people at a party. Some know each other (edge), others don't (no edge). Bill, Sue, Bob, and Jill all know each other; Sam, Sue, Rachel, and Sarah are mutual strangers.



# Soap Solution

Soap is slippery stuff. So, apparently, is the mathematical theory of soap bubbles, that is. When it comes to the geometric properties of soap bubbles, there are more questions than answers. And even when an answer seems firmly in hand, the proof can be as hard to get hold of as—well, as a wet bar of soap.

Chief among the unsolved problems: What shape or shapes will a cluster of soap bubbles assume? It's well known that a *single* bubble minimizes its surface area for the volume it contains by assuming the shape of a sphere. But what happens when *two* bubbles get together? It sounds like a straightforward problem in geometry, with a little bit of calculus thrown in. Surprisingly, the answer is still not known. Or rather, the answer is thought to be known, but so far there is no proof that the answer is correct.

There has been progress, however. For the last several summers, groups of students in a summer Research Experience for Undergraduates (REU) program at Williams College in Williamstown, Massachusetts, have gotten their hands dirty with the theory of soap bubbles. Working with faculty advisor Frank Morgan, an expert in geometric measure theory, the students have taken on—and solved—some subtle problems in the geometry of soap. One group's results appeared in 1993 as a paper in the *Pacific Journal of Mathematics*; other papers are in the pipeline.

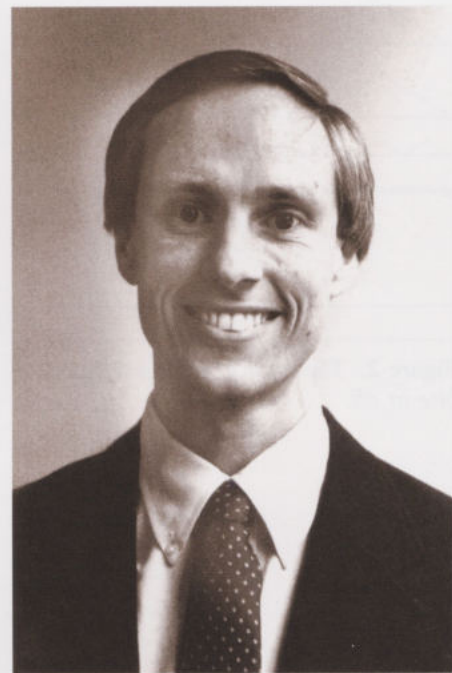
The Williams College REU is one of several dozen summer programs around the country offering students a chance to work on open problems in mathematics. More than a hundred students have participated in the Williams program since 1988, working on problems in the theories of knots, numbers, and graphs, along with the geometry of soap bubbles.

Working in the Williams REU “made me look at mathematics differently,” says Joel Foisy, who is now a graduate student at Duke University. “Even now it's helping me, because I know that I'm capable of doing original research.”

Jeff Brock, now in graduate school at the University of California at Berkeley, agrees. The summer experience was “a definite turning point” in his career. “We were given a tremendous amount of freedom to go about things as we liked,” Brock recalls. “When things were going well, I could spend hours at a time thinking about it. And it was really exciting when we actually started getting results.”

They had good reason to get excited. In the research leading to their *Pacific Journal* paper, carried out in the summer of 1990, Foisy and Brock, along with Manual Alfaro, Nickelous Hodges, and Jason Zimba, solved the 2-dimensional “double bubble” problem: Given two prescribed areas, find a pair of shapes in the plane whose combined perimeter is as small as possible. In other words, suppose you want to build a pair of corrals of particular sizes (say 1 acre for your sheep and 2 acres for your horse). How should you lay out the corrals to use as little fence as possible?

The suspected answer was the so-called “standard” double bubble (see Figure 1). The two bubbles are separated from the rest of the plane and from each other by circular arcs which meet at angles of 120 degrees (if the two bubbles are of equal size, then the inner arc is a straight line segment). The students showed that every other kind of double bubble has greater perimeter.



Frank Morgan. (Photo courtesy of William Tague, 1989.)

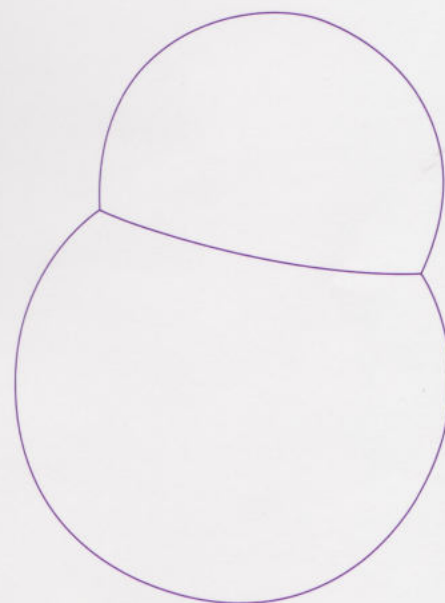
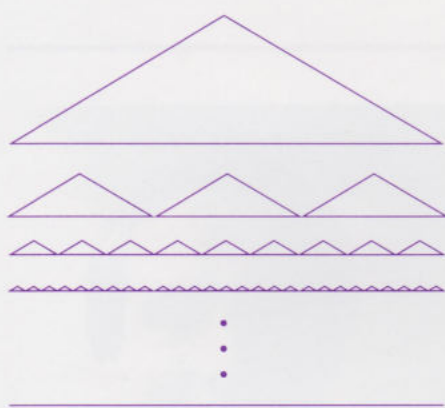


Figure 1. A standard double bubble. (Figures 1–5 courtesy of Silvio Levy, Geometry Center, Minneapolis, Minnesota.)





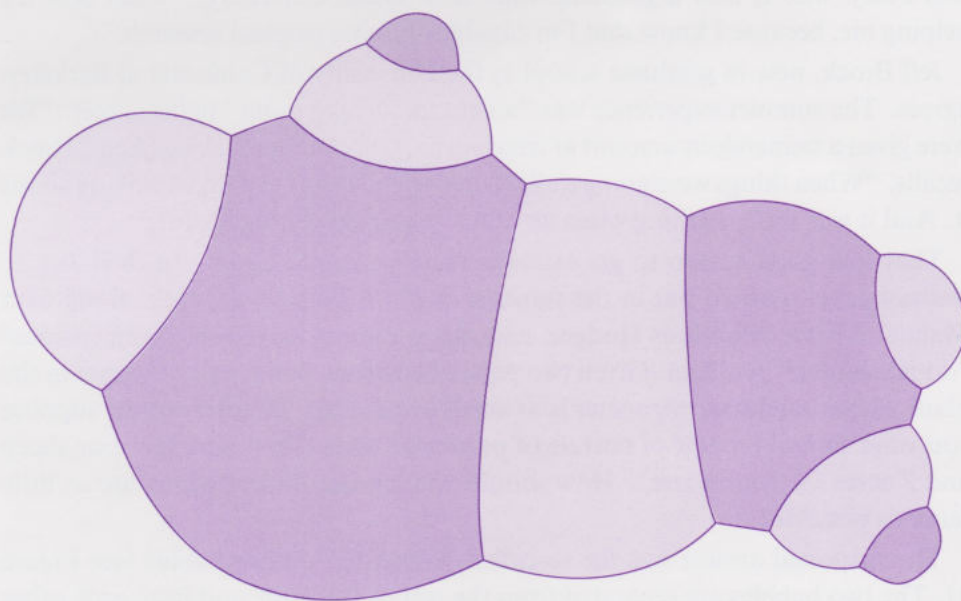
**Figure 2.** *The limit of sawteeth has no bite at all.*

It may seem surprising that this problem hadn't been solved long ago. After all, the single-bubble version, which asserts that the circle is the shortest curve enclosing a given area in the plane, was solved nearly 300 years ago with the advent of the calculus of variations, which treats functions, rather than numbers, as variables. (The problem itself dates back to antiquity. According to legend, when Queen Dido founded Carthage, she "merely" asked for as much land as she could contain within the hide of a bull. She then cut the hide into thin strips, fashioning an enormously long belt, which encompassed a sizable area. Whether Dido actually knew about the area-maximizing property of the circle is unclear, but Carthage was a major power for many years.)

So what held up the double bubble?

For one thing, it wasn't even certain that a solution existed. Conceivably, there could be a sequence of increasingly complicated bubble arrangements, each with less perimeter than its predecessor, but not converging to any definite final form. Such problems are perennial in the calculus of variations. For example, among sawtooth curves of fixed length, there is none that minimizes the area beneath the curve (see Figure 2): The area decreases as the teeth get finer, but the "limit" has no teeth at all!

For bubbles in 3 dimensions (and higher), only in the last 20 years have researchers nailed down the existence part of the theory. In the mid-1970s, Fred Almgren at Princeton University introduced a new geometric definition for soap bubbles and proved that solutions exist to problems of separating specified volumes with minimal surface area. Using Almgren's results, Jean Taylor at Rutgers University proved that these mathematical solutions had the "right" properties: Surfaces meet in threes at angles of exactly 120 degrees, and the seams they form meet four at a time at angles of approximately 109 degrees. These "regularity" properties had been observed in real soap bubbles over 100 years ago, but Taylor was the first to prove that no other behavior is possible.

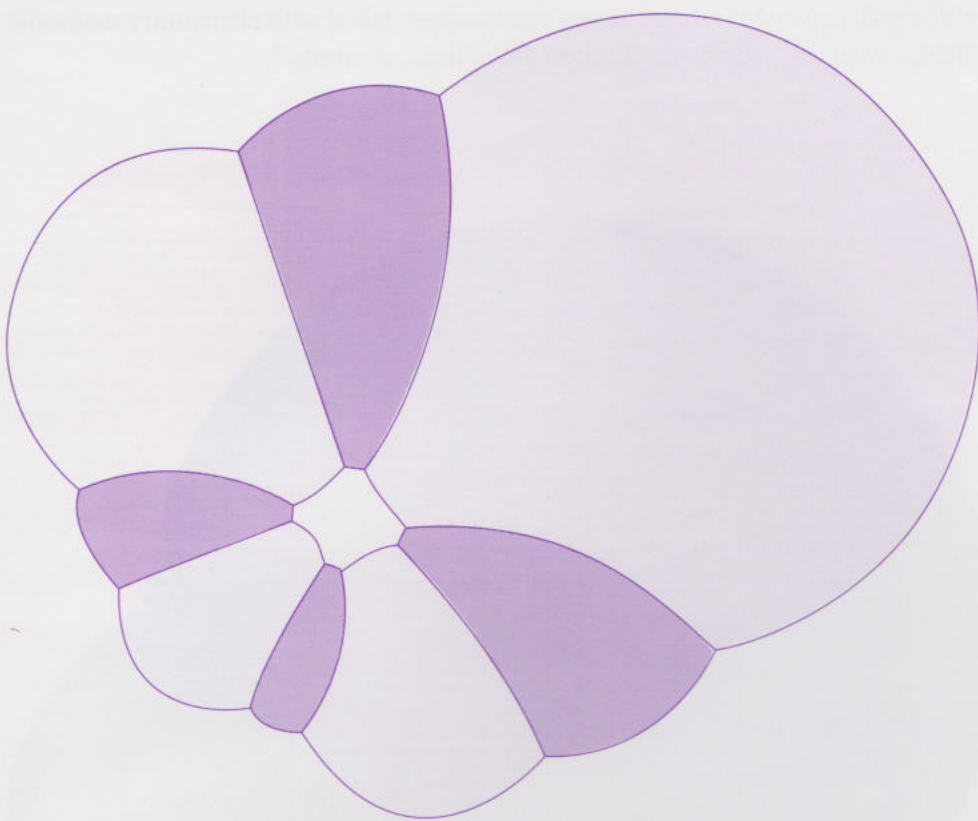


**Figure 3.** *Could the best double bubble have more than two components? Not in the plane, it can't.*



More recently, Morgan has proved analogous existence and regularity results for 2-dimensional bubbles, both for the plane and for more general (curved) surfaces. Morgan's results gave the students a theoretical basis from which to start.

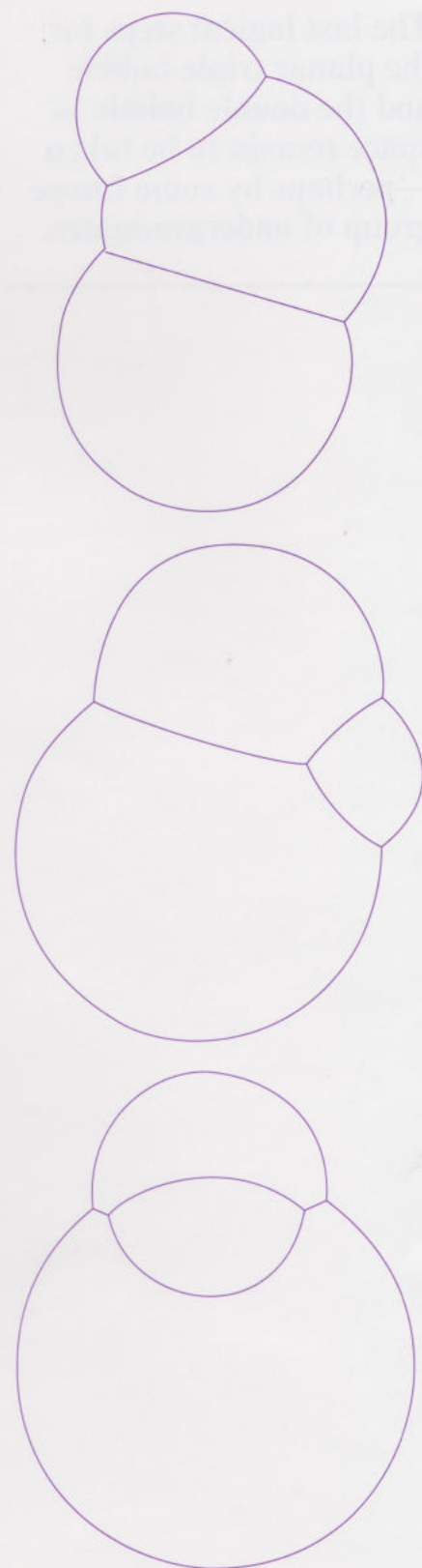
Even so, there was a lot of work left to be done. "The existence theory admits some very funny things," says Morgan. The main stumbling block is the theoretical possibility that the best way to minimize the total perimeter enclosing several prescribed areas (or the total surface area enclosing several prescribed volumes) is to split each area into several disconnected components. For example, the best "double" bubble might actually be a cluster with nine components, five of which comprise one area and the rest the other (see Figure 3). The existence theory even allows for the possibility that the exterior region has more than one component—that is, there might be "empty chambers" within a perimeter-minimizing bubble cluster (see Figure 4).



**Figure 4.** *Could the best double bubble have an empty chamber? Not in the plane, it can't.*

Foisy, Alfaro, Brock, Hodges, and Zimba eliminated those possibilities for the planar double bubble in the summer of 1990. (Brock was an undergraduate at Yale University at the time; the others were students at Williams.) The strategy of their proof had two parts. They first showed that if a perimeter-minimizing double bubble has no empty chambers, then it must be the expected "standard" double bubble; they then showed that empty chambers cannot occur.

In 1992 another group of students took on the 2-dimensional *triple* bubble problem. Chris Cox, Lisa Harrison, Michael Hutchings, Susan Kim, Janette Light, Andrew Mauer, and Meg Tilton showed that, if the perimeter-minimizing



**Figure 5.** *The standard triple bubble in the plane (center) encloses three specified areas with the least perimeter. Two other types of triple bubbles (top and bottom) do not do as well.*



---

**The last logical steps for the planar triple bubble and the double bubble in space remain to be taken—perhaps by some future group of undergraduates.**

---

solution for three areas is assumed to consist of connected regions, then the “standard” form wins out over two other combinatorial possibilities (see Figure 5). More recently, Hutchings, now a graduate student at Harvard University, has gone back to the double bubble—but up a dimension. It’s known that the surface-area-minimizing double bubble is a “surface of revolution” obtained by rotating some plane curve around an axis. The likely answer is the shape that results when the 2-dimensional standard double bubble is spun around its axis of symmetry, but a proof remains elusive. However, Hutchings has shown that the solution, whatever it may look like, has no empty chambers. Moreover, in special cases, such as when the two prescribed volumes are equal, he has proved that the enclosed regions are also connected.

The last logical steps for the planar triple bubble and the double bubble in space remain to be taken—perhaps by some future group of undergraduates. The problems are good ones for students, says Hutchings, because the high-level part of the problem is done, and the rest can be approached with elementary methods: “It just requires some determination and a little ingenuity.”



---

**Figure 6.** *A 3-dimensional double bubble.*

---



# Straightening Out Nonlinear Codes

**E**ver since computers started taking over the bulk of the work in data processing and telecommunications, people who use them have worried over a fundamental question: How do you cope when the machine malfunctions?

In the beginning, computers and the programs they ran were simple enough that physical failures—usually the death of a vacuum tube—were readily apparent. But as hardware advanced and programs got more elaborate, the prospect of microscopic flaws that alter how a machine runs or how it handles data became a real concern. With chips getting smaller every year, and computers getting faster, the chance of an occasional error slipping in gets better and better. Even when that chance is one in a billion, a computer running at 25 megaHertz—the speed of last year's laptops, ponderously slow by supercomputing standards—is going to screw up 90 times an hour. What's to be done?

The answer, researchers found, lay in what are known as error-correcting codes. The mathematical theory of these codes, developed over the last 40 years, has enabled computer scientists and engineers to design systems that work reliably at the very edge of their physical capabilities. Error-correcting code technology is nowadays as common as compact disks; it's what allows your favorite Mozart or Madonna CD to play perfectly even though your cat's been clawing the disk. The same technology has been used in deep-space probes, allowing spacecraft such as *Voyager II* to send back sparkling-clear pictures of distant planets while using less power than a refrigerator lightbulb.

---

**Error-correcting code technology is nowadays as common as compact disks; it's what allows your favorite Mozart or Madonna CD to play perfectly even though your cat's been clawing the disk.**

---



*Left to right: Roger Hammons, Jr., Patrick Solé, Vijay Kumar, Robert Calderbank, and Neil Sloane. (Photo courtesy of Elizabeth Cruger Arts Photography.)*

Error-correcting technology may soon get even better, thanks to some new discoveries in coding theory. Two separate groups of researchers recently found the key to a set of powerful error-correcting codes that have the awkward property of being "nonlinear." Roger Hammons, Jr., at Hughes Network Systems in Germantown, Maryland, and Vijay Kumar at the University of Southern California teamed up with Robert Calderbank and Neil Sloane at AT&T Bell Laboratories



---

**Coding theorists have known for decades that nonlinear codes of a given length can have more code words than their linear counterparts.**

---

in Murray Hill, New Jersey, and Patrick Solé at the Centre National Recherches Scientifique in Sophia Antipolis, France, to show that many nonlinear codes can actually be considered linear—when looked at in the right way.

Their findings, which will appear in the *IEEE Transactions in Information Theory*, “open the gates” to the use of nonlinear codes, says Kumar. Among the promising applications is “sequence design” for digital cellular communications, which will eventually replace the analog technology now used in gadgets such as car phones. Systems based on nonlinear codes could serve many more users with the available bandwidths.

But first of all, what are error-correcting codes, and what does linearity have to do with the subject?

In general, a mathematical code is simply a set of “words,” each of which is nothing more than a string of symbols. The most commonly used “alphabet” has just two symbols: 0 and 1. Typically, the words in a code all have the same “length”—that is, the same number of 0s and 1s. (Not all codes work that way. The familiar Morse code, for example, uses a simple *dot* to represent the frequently appearing “word” *e*, but a longer *dot-dot-dash-dot* for the less common *f*.)

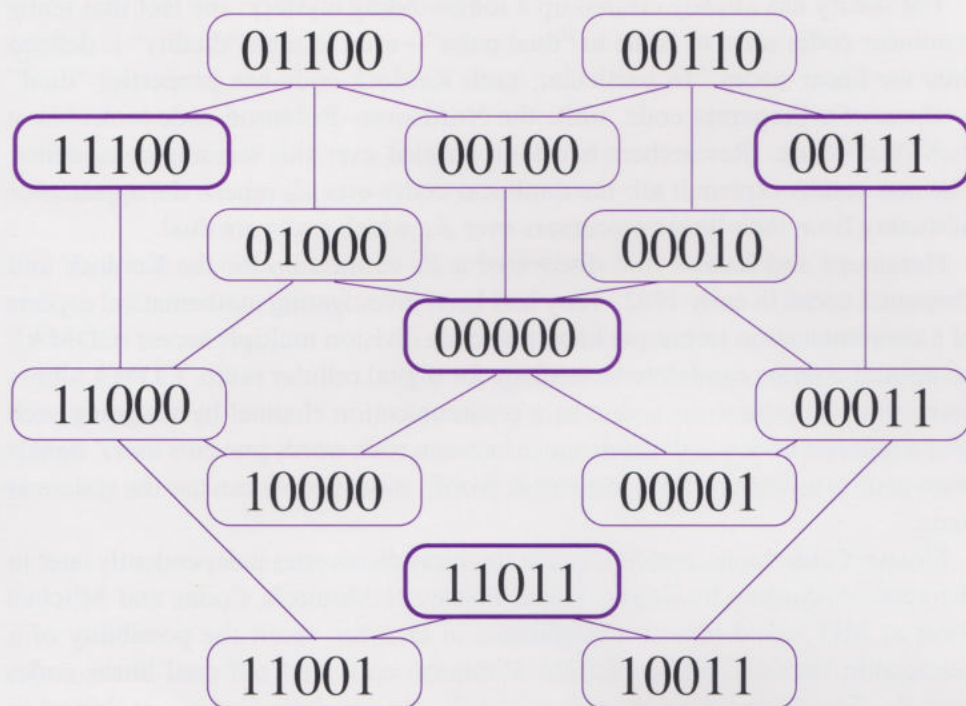
The error-correcting capability of a code is based on a notion of “distance” between code words. The distance between two words is simply the number of places in which they differ. If the distance between any two code words is at least 3, then the code can correct single errors. For example, in the code {00000, 11100, 00111, 11011}, the misread word 01100 can be corrected to its “nearest neighbor” 11100, from which it differs in only one digit (it differs from the other code words in two, three, and four places, respectively). Likewise, when the distance between code words is at least 5, the code can correct double errors; triple error-correcting requires distance 7 or more, and so forth.

The code {00000, 11100, 00111, 11011} is also an example of a *linear* code: If you add two code words together using the binary addition rule  $1 + 1 = 0$ , the result is another code word. For example,  $11100 + 11011 = 00111$ . Linearity gives a code an algebraic structure that makes decoding messages much easier and makes encoding them a snap. In precise mathematical terms, a linear code is a vector space over the finite field  $Z_2$ —so the full force of linear algebra can be brought to bear.

But linear codes also have their downside. The main problem is that linearity often restricts the number of possible code words, which hinders the code’s ability to carry information. If you’re not worried about linearity, you can toss in as a new code word any string that maintains the appropriate distance from everything already in the code. But if you insist on linearity, then you have to check these distances not just for the prospective new code word, but also for all its sums with words already in the code.

Coding theorists have known for decades that nonlinear codes of a given length can have more code words than their linear counterparts. Among linear codes of length 16, for example, the best double-error-correcting code has 128 code words. But in 1967, A. W. Nordstrom and John Robinson constructed a nonlinear code containing 256 words (Nordstrom was a high-school student at the time, Robinson an electrical engineer at the University of Iowa). In effect, only 7 digits in each code word of the linear code carry information (the other 9 do the error correcting), whereas in the Nordstrom–Robinson code, 8 digits carry information, an improvement of approximately 14%. Researchers have developed many other





**Figure 1.** The code words {00000, 11100, 00111, 11011} and some of their neighbors.

examples of nonlinear codes, including two families of codes known as Kerdock and Preparata codes, both of which generalize the Nordstrom–Robinson code.

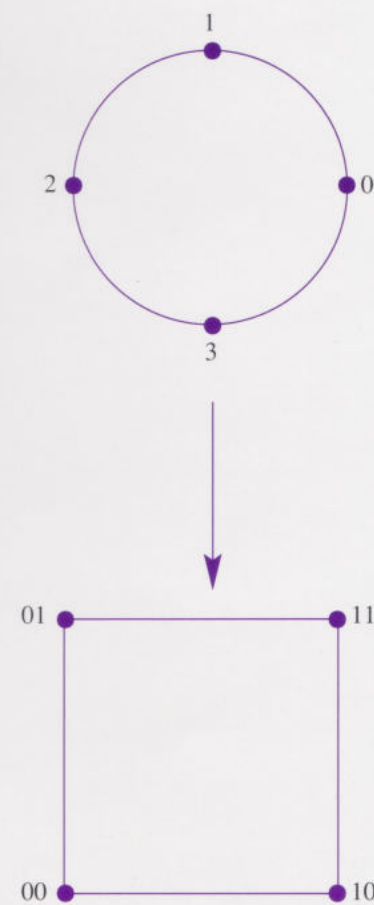
Even so, linear codes have predominated in practice, because their extra structure makes them easier to work with, which translates into faster and more efficient algorithms for encoding and decoding. Nonlinear codes' apparent lack of structure has left them in the dust.

Until now.

The five researchers have discovered a simple trick that turns many familiar nonlinear codes into linear codes—not over  $\mathbb{Z}_2$ , though, but over  $\mathbb{Z}_4$ : the algebraic system  $\{0, 1, 2, 3\}$  with the rules  $2+2 = 1+3 = 2 \times 2 = 0$  and  $3+3 = 2 \times 3 = 2$ . More precisely, they have found that many nonlinear codes can be obtained as “images” of codes that are linear over  $\mathbb{Z}_4$  using a particularly simple mapping.

The trick is a kind of squaring of the circle (see Figure 2). The system  $\mathbb{Z}_4$  is often represented by four points of a compass: 0 and 2 at East and West, 1 and 3 at North and South. Adjacent digits are considered to differ by 1, opposite digits by 2: the distance between code words reflects these differences. For example, the distance between 0000 and 0123 is 4, since the digits differ by 0, 1, 2, and 1, respectively. The new idea is to take a linear code over  $\mathbb{Z}_4$  and replace the digits 0, 1, 2, and 3 with 00, 01, 11, and 10, respectively. The result is a nonlinear code over  $\mathbb{Z}_2$  that is really just a linear code over  $\mathbb{Z}_4$  in disguise!

The surprise is that this trick accounts for essentially all of the nonlinear codes that theorists have studied so far, including the Nordstrom–Robinson, Kerdock, and Preparata codes. It didn't have to work out that way; the trick might have produced only a limited subclass of nonlinear codes—and uninteresting, useless ones, at that. The theory's success hints at deep connections among the various kinds of codes, with more surprises possibly in store.



**Figure 2.** The key to nonlinear codes: “squaring the circle.”



---

**The theory's success hints at deep connections among the various kinds of codes, with more surprises possibly in store.**

---

The theory has already cleared up a longstanding mystery: the fact that many nonlinear codes seem to come in “dual pairs”—even though “duality” is defined only for linear codes. In particular, each Kerdock code has properties “dual” to those of a Preparata code, while the Nordstrom–Robinson code looks like a “self-dual” code. Researchers had long puzzled over this seeming coincidence. The new results explain it all: the nonlinear codes over  $Z_2$  inherit the appearance of duality from their linear precursors over  $Z_4$ , which really are dual.

Hammons and Kumar first discovered a  $Z_4$  connection for the Kerdock and Preparata codes in early 1992. They had been investigating mathematical aspects of a communication technique known as code division multiple access (CDMA), an up-and-coming candidate technology for digital cellular radio. CDMA allows many users simultaneous access to a communication channel by assigning each user a separate code word; the distance between code words prevents users' signals from getting mixed up. With more code words, more people can use the system at once.

Sloane, Calderbank, and Solé made the same discoveries independently later in the year. According to Sloane, David Forney at Motorola Codex and Mitchell Trott at MIT asked him at a conference in October about the possibility of a connection between the Nordstrom–Robinson code and self-dual linear codes over  $Z_4$ . Sloane was the right person to ask. He and John Conway at Princeton University had recently completed a study of such codes, so he immediately knew which code would give the connection, if there was one: a self-dual code over  $Z_4$  known as the octacode.

“I went home and, in two minutes thinking about it, it became clear that yes indeed, the octacode was really the same thing as the Nordstrom–Robinson code,” Sloane recalls. “I called up Conway and said, ‘Look! How could we have missed this? We should have noticed this years ago!’”

Calderbank and Solé contributed several key ideas to Sloane's observation, and the three of them soon had an extensive theory, including efficient algorithms for decoding the Kerdock and Preparata codes. Then Calderbank discovered that Hammons and Kumar had found many of the same results. The two groups agreed to publish their results jointly.

More recently, Sloane and Conway have found  $Z_4$  precursors for a number of single-error-correcting codes that are nonlinear over  $Z_2$ , while Calderbank, Kumar, and Tor Hellesest at the University of Bergen, in Norway, have discovered some new codes over  $Z_4$  which are better, by various technical standards, than any of the previously known families over  $Z_2$ . Together with Peter Cameron at Queen Mary and Westfield College in London, Bill Kantor at the University of Oregon, and Jaap Seidel at the Technical University of Eindhoven in the Netherlands, Calderbank has also begun investigating a surprising connection between the  $Z_4$ -linearity of the Kerdock codes and some seemingly unrelated problems in finite geometry. Nonlinear codes may finally be getting straightened out, but it looks like coding theorists can still count on quite a few twists and turns.



# Quite Easily Done

**T**he line between easy mathematical problems and hard ones is finely drawn. Some problems seem to cross back and forth: First they look easy, then they seem hard, and then, when they're finally solved, they look easy again. A recent example is a simple-sounding combinatorial puzzler called the Dinitz problem. First posed in 1978, the Dinitz problem has finally been solved with a surprisingly simple proof, but only after fifteen years during which it seemed a very tough nut to crack.

The story starts in the late 1970s. Jeff Dinitz, then a graduate student at Ohio State University (now a professor at the University of Vermont), was studying properties of combinatorial arrangements known as latin squares. A latin square is an  $n \times n$  array of  $n$  symbols—say a  $5 \times 5$  array of stars, squares, circles, diamonds, and triangles—in which no symbol appears more than once in any row or column (see Figure 1). Latin squares are useful, for example, in the design of experiments, to protect against bias. If, say, you want to compare five different herbicides in a corn field, but want to make sure the results aren't affected by variations in soil quality from one side of the field to another, then dividing the field into a  $5 \times 5$  latin square pattern is an efficient way to design the experiment.

Latin squares are easy to come by. Indeed, their number explodes with the size of the square, from two  $2 \times 2$  squares to twelve  $3 \times 3$  squares to more than  $10^{19}$  squares of size  $8 \times 8$ . But Dinitz cooked up a variant on the problem of constructing latin squares for which it wasn't clear—until now—that *any* solution could be found.

In an ordinary  $n \times n$  latin square, there is only one set of  $n$  symbols, and an element from that set must be chosen for each location in the square. In Dinitz's version—called a “partial latin square”—each location is assigned its *own* set of  $n$  possible symbols: these sets may vary from location to location. The problem is still to choose a symbol for each location, but now the symbol must come from the set assigned to that location. The goal, however, remains the same: to avoid choosing the same symbol twice in any one row or column.

In Figure 2, a three-element set is assigned to each location in a  $3 \times 3$  square; the elements in orange constitute a partial latin square. The Dinitz problem asks: Given any assignment of  $n$ -element sets of symbols to the  $n^2$  locations in an  $n \times n$  array, is it always possible to find a partial latin square? Or to put it negatively, among all the ways to assign  $n$ -element sets to the locations of an  $n \times n$  array, are there any for which it's impossible to pick an element from each set without picking some symbol twice in the same row or column?

At first glance, the answer seems obvious: Since the problem, in general, uses more than  $n$  symbols, it should be easier to satisfy the nonrepetition requirement for a partial latin square than for an ordinary latin square. But that glance overlooks a crucial aspect of the problem: Not every symbol is available at every location. One way to construct an ordinary latin square is to specify where in each row you'll place the first symbol, where the second symbol, and so on; that approach doesn't even make sense for partial latin squares.

Another telling difference between ordinary and partial latin squares casts further doubt on the “obviousness” of the answer. Ordinary latin squares can always be filled in “row by row.” If, say, the first two rows of a  $5 \times 5$  square have been



**Figure 1.** Each of five symbols appears exactly once in a  $5 \times 5$  latin square.



**Figure 2.** One symbol (orange) from each three-element set can always be chosen to form a  $3 \times 3$  partial latin square.





Jeff Dinitz

filled in successfully (without doubling up in either row or any column), then the rest of the rows can also be filled in to give a latin square. That means that when you're trying to create a latin square, you'll never paint yourself into a corner—you won't get down to the last row, for example, and find yourself unable to complete the square. With partial latin squares, by contrast, you *can* paint yourself in. For example, if the sets in the first row of a  $2 \times 2$  array are  $\{A, B\}$  and  $\{B, C\}$ , it's natural to choose  $A$  and  $B$  as the symbols in that row—but then you get in trouble when you see the sets  $\{A, C\}$  and  $\{B, C\}$  in the next row.

Complications notwithstanding, Dinitz's conjecture—that partial latin squares can always be found—turns out to be true. It just took fifteen years for a proof to be found. In the meantime, the problem served as a kind of drawing card for the theory of combinatorial design and a testing ground for new ideas.

Dinitz's conjecture can be verified directly for  $2 \times 2$  arrays, because there are so few different possibilities. In principle, the conjecture can be checked for arrays of any given size. That's because there are only finitely many cases to check: The total number of distinct symbols for an  $n \times n$  array cannot exceed  $n^3$ , so the number of cases is less than  $n^3$  to the power  $n^3$  (more precisely, it's at most the  $n^2$  power of  $\binom{n^3}{n}$ ). But the numbers involved in such a brute-force, case-by-case analysis grow astronomically with  $n$ . The  $3 \times 3$  problem is small enough for this approach to be practical, but the  $4 \times 4$  case is already out among the stars.

In 1991, however, Noga Alon and Michael Tarsi at Tel Aviv University in Israel proved a theorem that made it easy to verify (by computer) Dinitz's conjecture for  $4 \times 4$  and  $6 \times 6$  arrays. Their theorem is not specific to Dinitz's problem. It concerns a general problem in graph theory called "list coloring."

In combinatorics, a graph is a set of points (called *vertices*) and a set of lines or curves (called *edges*) connecting them. Many applications of graphs in scheduling or network theory can be interpreted as coloring the edges of a graph, with the stipulation that no two edges of the same color meet at a common vertex. To schedule a college football season, for example, let each team be represented by a vertex, draw an edge connecting teams that are slated to meet, and then color each edge according to the week on which the two teams are to play (say red for week 1, blue for week 2, and so on). The condition that no like-colored edges should meet at a common vertex simply means that no team should be asked to play two games simultaneously.

In a list-coloring problem, each edge in a graph is assigned a prescribed set, or list, of allowed colors. The Dinitz problem can be viewed as a special case of list coloring, for graphs in which each of  $n$  "row" vertices is joined to each of  $n$  "column" vertices (see Figure 3). Graphs of this type, in which the vertices are separated into two sets and all edges cross from one set to the other, are known as "bipartite" graphs; the particular graph associated with the Dinitz problem is called a complete bipartite graph, because it includes all possible edges between the two sets of vertices. There is a general conjecture regarding how large the palette of possible colors for each edge of a graph must be in order to ensure that a list coloring is possible. Viewed from this angle, the Dinitz problem is just the tip of an immense theoretical iceberg.

Alon and Tarsi's theorem gives a condition which, if satisfied, guarantees the existence of a list coloring from sets of a particular size. Their condition is simple enough to be verified explicitly for the graphs associated with the  $4 \times 4$  and  $6 \times 6$  Dinitz problems. In principle, the condition can be checked for *all* even  $n$ ,



but once again, the amount of computation involved gets quickly out of hand. Furthermore, the condition is *never* satisfied for odd  $n$ . (This doesn't mean that the Dinitz conjecture is false for odd  $n$ , just that Alon and Tarsi's theorem won't help prove it for those cases.)

Other researchers, notably Roland Häggkvist at the University of Stockholm, had made inroads on the list coloring problem and its relation with the Dinitz conjecture. In late 1992, Jeannette Janssen, then a graduate student at Lehigh University in Bethlehem, Pennsylvania (now a postdoc at Concordia University in Montreal), proved a result that surprised even many of the experts. Janssen showed that Alon and Tarsi's theorem could be used to solve completely a slightly weaker version of Dinitz's problem. Instead of focusing on squares, Janssen looked at *rectangles*—arrays with fewer rows than columns. She showed that in any  $r \times n$  array with  $r < n$ , it's enough to have  $n$  symbols (or colors) assigned to each location in order to guarantee that a partial latin rectangle exists.

Janssen's result comes close to the full Dinitz conjecture in two different (but closely related) ways. First, it says that you can always fill in at least the first  $n - 1$  rows of a partial latin square (the previous best result guaranteed only two-sevenths of the rows). Second, by starting with an  $n \times (n + 1)$  rectangle and then lopping off the last column, Janssen's theorem says that you can always find a



Jeannette Janssen. (Photo courtesy of Cliff Skarstedt.)

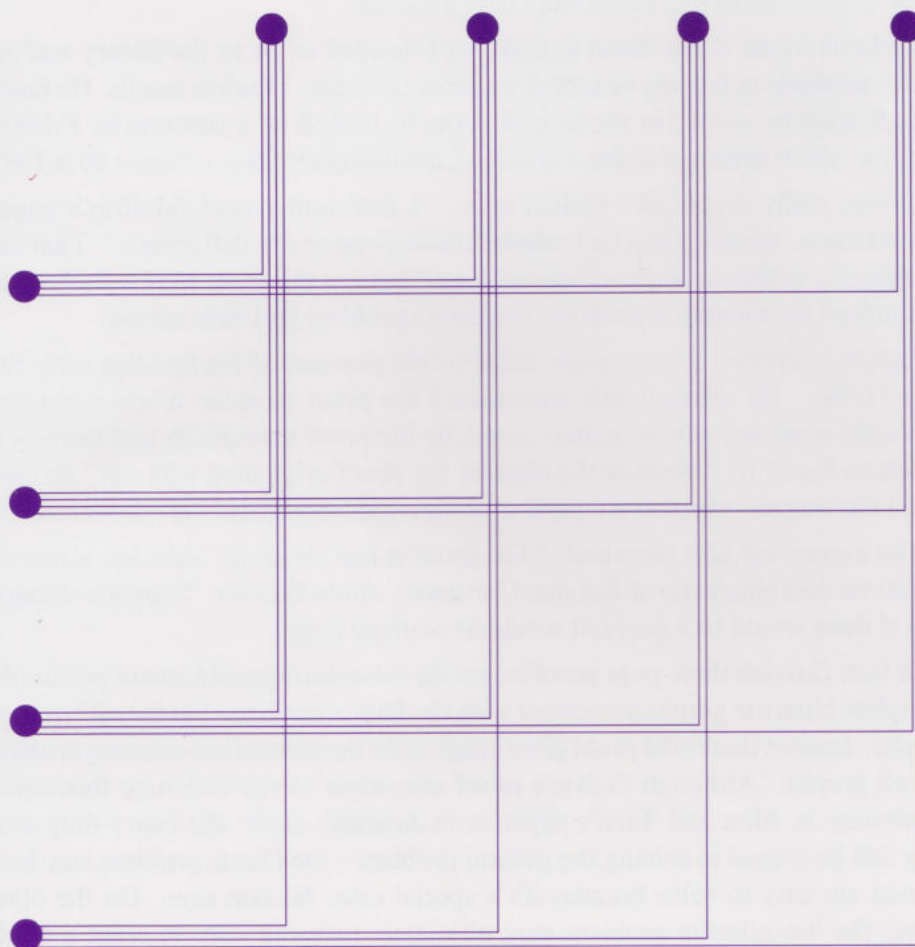
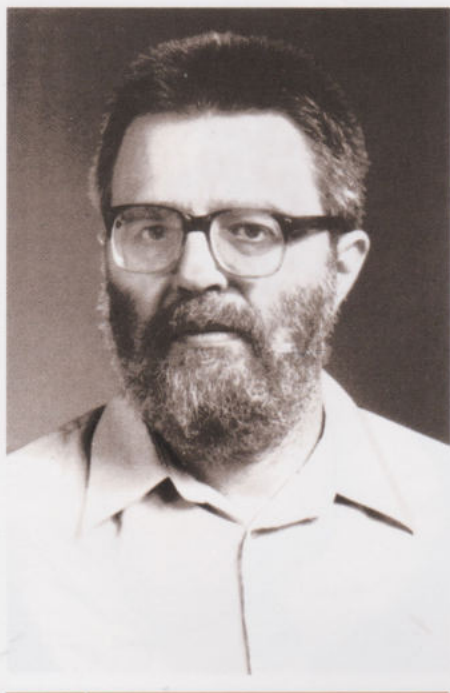


Figure 3. Each edge in a bipartite graph corresponds to a location in a  $n \times n$  array.





Fred Galvin

partial latin square if  $n + 1$  symbols have been assigned to each location—again, far better than previous results.

Experts in the field lauded Janssen's breakthrough. "It is brilliant," said Herb Wilf of the University of Pennsylvania. "It moves the problem much closer to a resolution than anyone had expected." Other theorists agreed, predicting the full Dinitz problem would be solved soon, perhaps within a year. They were right—but not quite for the reasons they had in mind.

Fred Galvin, a mathematician at the University of Kansas, read Janssen's proof in the *Bulletin of the American Mathematical Society*; this led him back to Alon and Tarsi's paper in the journal *Combinatorica*. A remark in that paper made Galvin realize that one of the ideas in Janssen's work could be parlayed into a proof of the complete Dinitz problem, provided one could prove a certain result about the existence of something called a kernel.

Loosely speaking, a kernel of a graph is a "largest possible" subset of vertices, no two of which are connected by an edge. The precise definition is more technical, but the way kernels are used in Galvin's proof is simple: take any color, say red, identify the set of locations that include red among their allowed colors, find a kernel of that set, and then make red your choice for all the locations in that kernel. The Dinitz problem is solved by repeating this process with other colors until every location has been colored—but this approach wouldn't work, Galvin knew, if some set of locations didn't have a kernel.

"I didn't know much about kernels, so I decided to go to the library and see what's available in the way of kernel existence theorems," Galvin recalls. He found exactly what he needed in the second paper he looked at, a theorem by Frédéric Maffray which appeared in the *Journal of Combinatorial Theory (Series B)* in 1992.

"I was really surprised," Galvin says. "I read and reread [Maffray's paper] several times, thinking maybe I misunderstood one of the definitions." That can happen in a technical tangle of terminology—but not this time. Maffray's theorem was indeed the missing ingredient; the Dinitz problem had been solved.

Galvin circulated a three-page, handwritten account of his findings early this year (1994). He subsequently streamlined the proof to make it self-contained. He is still surprised, almost embarrassed, by the proof's simplicity and the way in which he found it. "None of the ideas in the proof originated with me," he says. "All I did was put together a couple of things that were already in the literature."

The experts are also surprised. "The proof is just amazing," says Jeff Kahn, an expert on combinatorics at Rutgers University. Adds Janssen: "Nobody thought that if there would be a proof, it would fit on three pages."

In fact, Galvin's three-page proof solves the list-coloring problem not just for the complete bipartite graphs associated with the Dinitz problem, but for *all* bipartite graphs. Janssen thinks the proof gives insight into the general list-coloring problem for all graphs. Although Galvin's proof uses none of the elaborate theoretical machinery in Alon and Tarsi's paper or in Janssen's work, the heavy-duty stuff may still be crucial in solving the general problem—the Dinitz problem may have turned out easy to solve because it's a special case, Janssen says. On the other hand, the list-coloring problem may ultimately turn out easy to solve as well, perhaps because it's a special case of some even more general problem. If there's a lesson to be drawn, it's that hard problems need not stay that way.



---

**Computer science is rife with problems for which solutions indisputably exist, but for which efficient algorithms to find them are lacking.**

---

## The Road Least Traveled

Fred Galvin's solution of the Dinitz problem (see main story) not only shows that partial latin squares exist, it also points to an efficient algorithm for finding them. That doesn't always happen: Computer science is rife with problems for which solutions indisputably exist, but for which efficient algorithms to find them are lacking.

In the classic example, known as the Traveling Salesman Problem, a salesman (or woman) starts at the home office, visits a certain set of cities, and returns home. The "cost" (in time, mileage, or money) of traveling between each pair of cities is known: the objective is to complete the calls at the least possible total cost.

The Traveling Salesman Problem is an example of a "combinatorial optimization" problem—"combinatorial" because it deals with ways of arranging things, and "optimization" because it asks for the best arrangement. Like the Dinitz problem, the Traveling Salesman Problem can be phrased in graph-theoretic terms: The vertices of the graph are the cities, and the edges are the roads connecting them.

The problem and its variants have a foot in the door of many applications in which resources need to be routed. The manufacture of printed circuit boards presents one example. In order to connect conductors on different layers of a printed circuit board, it's necessary to drill holes—as many as several thousand, nowadays. The job is best done by a robot, which never gets bored or takes coffee breaks. But even a robot can waste time. The drilling robot must pick up the right size drill bit, move from hole to hole, and then return the bit (perhaps to exchange it for one of a different size). Moving the drill about is necessary, but unproductive and time-consuming; ideally, the drill will move as little as possible.

In principle, solving the Traveling Salesman Problem is easy: If the salesman has  $n$  cities to visit (including the home office), then only  $(n - 1)!/2$  different routes are possible, so it's just a matter of checking to see which one is shortest (or cheapest). The catch, of course, is in that "only." The number of possible routes grows exponentially with the number of cities, making a brute-force approach impractical for any problem with more than a handful of cities.

Computer scientists draw the line at programs whose run time grows exponentially with the size of the problem. They much prefer "polynomial-time" algorithms: programs whose run time grows no faster than some power of the problem size (see "Random Algorithms Leave Little to Chance," pages 27–32). But so far no one has found a polynomial-time algorithm for solving the Traveling Salesman Problem. Indeed, the general consensus is that no such algorithm exists: solutions to the Traveling Salesman Problem, it's thought, are inherently hard to find, even though they obviously exist.

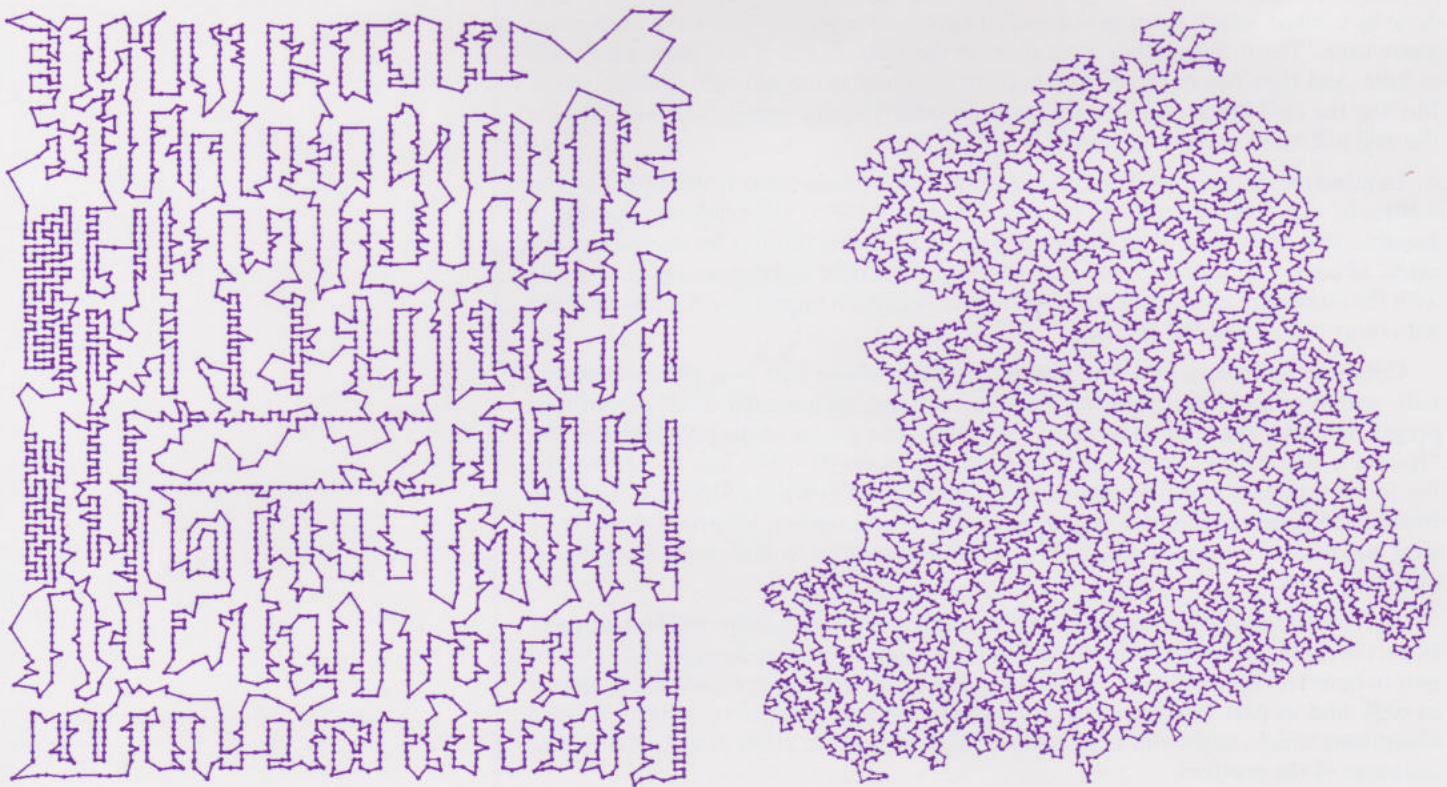
That hasn't kept people from looking for better ways to tackle the problem, though. In part because the Traveling Salesman Problem crops up repeatedly in applications, in part to hone techniques that can be used in other combinatorial optimization problems as well, and in part just because the challenge is there, researchers have developed algorithms which, while still exponential, manage to solve some exceptionally large instances of the problem.

David Applegate at AT&T Bell Laboratories, Bob Bixby at Rice University, Vasek Chvatal at Rutgers University, and Bill Cook at Bell Communications Research (Bellcore) in Morristown, New Jersey, have come up with what might be the best approach yet. Their algorithm stems from a method introduced in 1954, when computers—and combinatorial optimization—were just getting off the ground. The basic idea is to convert the original problem into a sequence of linear programming problems: solving them gives an increasing sequence of lower bounds for the cost of the salesman's cheapest route. Each individual linear programming problem is easy to solve; the catch is, it may require solving a huge number of them to get at the final answer.



To avoid getting lost in endless computation, Applegate and colleagues have added a “branch and bound” technique. Their algorithm periodically picks a pair of cities and divides the search for an optimal route into two branches: routes that visit the two chosen cities consecutively, and routes that don’t. The search along a particular branch is curtailed (bounded) if that branch offers nothing better than a route that’s already known.

The new method has already seen some success. Applegate and colleagues have used their branch-and-bound technique to solve more than a dozen longstanding “challenge” problems, including one with 3038 “cities” (see Figure 4 (left)). As it happens, their toughest computation to date is, in a sense, already out of date: One of the challenge problems was to find the shortest tour of all 4461 cities in the former East Germany. The branch-and-bound algorithm chased the problem down a total of 2929 branches before coming up with the answer (see Figure 4 (right)).



**Figure 4.** A traveling salesman's best route around a printed circuit board (left) and the former East Germany (right). (Figures courtesy of Bill Cook, Bellcore.)



# (Vector) Field of Dreams

**W**hat goes around, comes around, right? Not necessarily. In fact, in the realm of 3-dimensional topology, what goes around need never come back around. At least that's one way to describe a recent result of Krystyna Kuperberg.

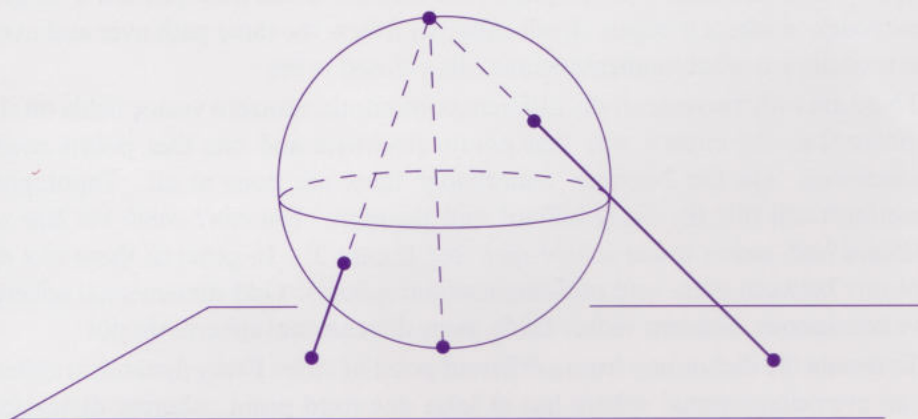
Kuperberg, a mathematician at Auburn University in Auburn, Alabama, has resolved a fortysome-year-old problem known as the Seifert conjecture. Dating back to a paper by Herbert Seifert in 1950, the Seifert conjecture concerns the topological properties of a 3-dimensional space, or manifold, known as the 3-sphere. A direct generalization of the ordinary circle and sphere (see box), the 3-sphere is, topologically, the simplest 3-dimensional manifold. But even so, many of its properties remain shrouded in mystery.

The Seifert conjecture, says John Franks, a mathematician at Northwestern University, "was the kind of question that we thought we should be able to answer—and we couldn't." Until now.

In technical terms, the Seifert conjecture asserts that every smooth, nonzero vector field on the 3-sphere necessarily has at least one closed orbit. This sounds eminently reasonable. Indeed, in his 1950 paper, Seifert proved that all vector fields of a certain class (namely, distortions of a vector field known as the Hopf



Krystyna Kuperberg.

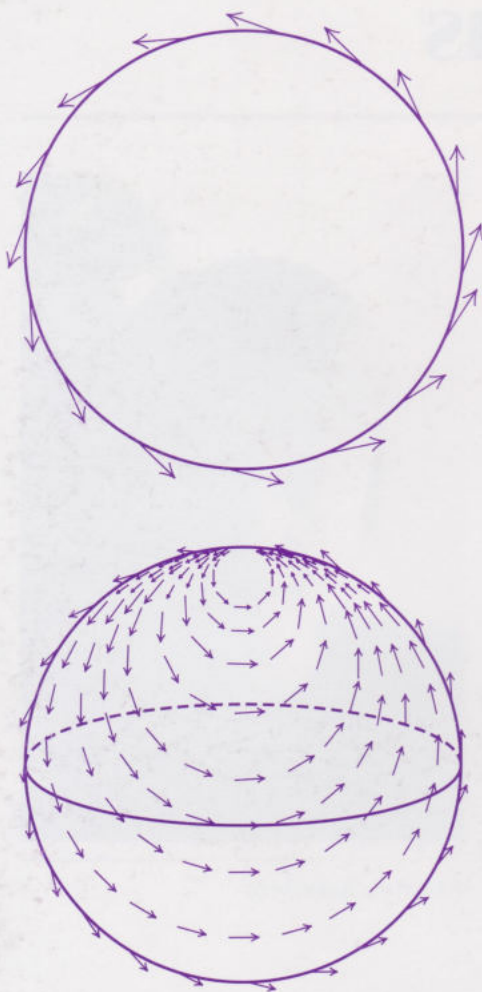


**Figure 1.** The "stereographic projection" maps every point in the plane to a point on the 2-sphere by connecting it to the "north pole," which can be thought of as corresponding to a "point at infinity" in the plane.

## Getting 'Round in $n$ Dimensions

The circle, known to topologists as the 1-sphere, or  $S^1$ , is the curve defined by the equation  $x^2 + y^2 = 1$  in the  $(x, y)$ -plane. Likewise, the "2-sphere"  $S^2$  is the surface defined by the equation  $x^2 + y^2 + z^2 = 1$  in 3-dimensional space. The  $n$ -sphere is a straightforward generalization: It is the  $n$ -dimensional "hypersurface" defined by the equation  $x_1^2 + x_2^2 + \dots + x_{n+1}^2 = 1$  in coordinates  $x_1, x_2, \dots, x_{n+1}$ . Topologically, the  $n$ -sphere is a compact version of  $n$ -dimensional Euclidean space with an extra "point at infinity." Figure 1 shows how the  $(x, y)$ -plane can be mapped onto the 2-sphere in 3-dimensional space. The corresponding picture in 4-dimensional space is left to the reader's imagination.





**Figure 2.** The 1-sphere (also known as the circle) allows for nonzero vector fields (top), but every vector field on the 2-sphere has a “bald spot.”

fibration) do have closed orbits. But even so, mathematicians soon came to doubt the general conjecture. Their doubts were well founded: The Seifert conjecture is false.

By adding an ingenious new twist to some old ideas, Kuperberg has constructed smooth vector fields with no closed orbits, thus putting the kibosh on Seifert’s conjecture—and not just for the 3-sphere, but for all 3-dimensional manifolds.

Kuperberg’s counterexamples could have implications in the theory of dynamical systems, where closed orbits correspond to periodic behavior, such as the regular swing of a pendulum or the predictable variations in a predator-prey relationship. Vector fields crop up constantly in the study of differential equations and mathematical physics. Newton’s law for gravitational motion and Maxwell’s equations for electromagnetism are just two examples where vector fields play a key mathematical role in describing physical phenomena.

Loosely speaking, a vector field assigns a little arrow to each point of the surface or space on which the field is defined. Arrows attached to different points can point in different directions, and they can have different lengths. The most familiar example of a vector field is wind: At each point on the surface of the earth, the wind can be described by an arrow pointing downwind, with length proportional to the windspeed. (Of course wind also changes in time. A vector field can be thought of as a wind that varies from place to place, but remains constant in time.) Through each point, a vector field determines a trajectory—the path a dust particle would follow if blown by the field’s wind. If the dust particle ever gets blown back to where it began, it will endlessly follow the same path over and over: The trajectory is what mathematicians call a closed orbit.

There are only two essentially different continuous, nonzero vector fields on the 1-sphere (i.e., the circle): one that points clockwise and one that points counterclockwise. On the 2-sphere, remarkably, there are none at all. Topologists sometimes call this the hairy billiard ball theorem: *You can’t comb the hair on a billiard ball, unless it has a bald spot* (see Figure 2). In general, there is a dichotomy between even- and odd-dimensional spheres: Odd-dimensional spheres have continuous, nonzero vector fields, even-dimensional spheres do not.

To restate the dichotomy from a different point of view: Every dynamical system on an even-dimensional sphere has at least one fixed point, whereas dynamical systems on odd-dimensional spheres need not have any fixed points. In effect, Seifert was asking whether dynamical systems on the 3-sphere have the next best property: Do those without fixed points necessarily have closed orbits?

Seifert’s original question referred generally to continuous vector fields, not specifically to smooth fields. (A vector field is “smooth” if the lengths and directions of the vectors change not just continuously, but smoothly—in technical terms, if the field is “infinitely differentiable.”) But in 1972, Paul Schweitzer at Pontificia University Católica in Rio de Janeiro, Brazil, produced a once-differentiable counterexample. A decade later, Jenny Harrison at the University of California at Berkeley constructed nonzero, orbitless vector fields that were twice differentiable. (Based on fractals, Harrison’s counterexamples could actually be differentiated up to—but not including—three times, given appropriate definitions for fractional differentiation.)

Schweitzer’s and Harrison’s once- and twice-differentiable counterexamples made Seifert’s conjecture seem unlikely to hold in the infinitely differentiable case either. On the other hand, there are plenty of theorems that hold for smooth



functions but lose their grip at any lesser level of differentiability. In any event, the constructions seemed stuck at the low-derivative end of things.

Kuperberg's construction breaks that impasse. Expanding on ideas she and Coke Reed, now at the Supercomputing Research Center in Bowie, Maryland, introduced in 1981 to resolve another conjecture about fixed points of dynamical systems, Kuperberg has shown how to modify a smooth vector field so as to break up any closed orbits that might be present. The construction is "very geometric," Kuperberg says. Keeping things smooth was not the hard part of the problem, she explains: "The main difficulty turned out to be not to form additional circular trajectories" in the process of modifying the field.

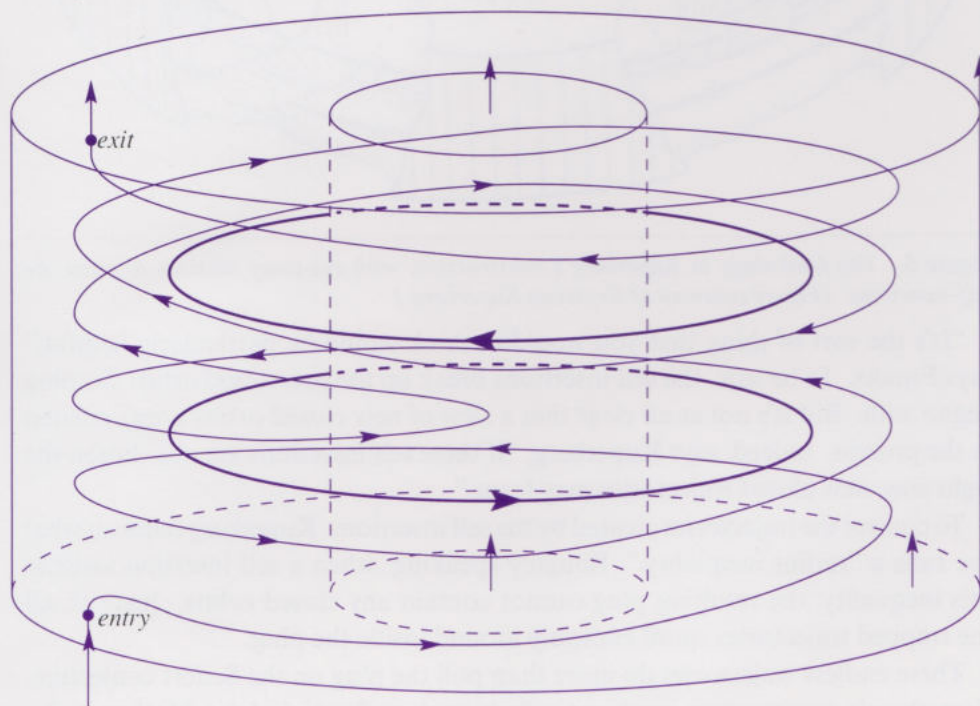
The starting point for Kuperberg's counterexample is a smooth vector field with finitely many closed orbits. (It's well known that such fields exist, even though vector fields with infinitely many closed orbits are easier to come by. Schweitzer and Harrison used the same starting point.) The basic tool in Kuperberg's construction is a topological gadget known as a "Wilson plug"—a 3-dimensional shape with a vector field that is constant on its boundary and which "traps" at least one trajectory that enters it. The idea is to pick a point on one of the closed orbits, look at a small neighborhood of that point using a coordinate system in which the vector field is constant, and then replace a piece of that neighborhood with the plug, arranging things so that the formerly closed orbit becomes one of the trajectories that enters the plug and gets trapped inside. The trick is to do this without creating any *new* closed orbits.

Kuperberg pulls off the trick in three steps. Curiously, in the first step she constructs a plug that has *two* closed orbits. At this stage the plug looks like a thick washer (see Figure 3). The vector field points straight up on the boundary, but inside the plug, the vectors change direction: Trajectories are deflected counter-

---

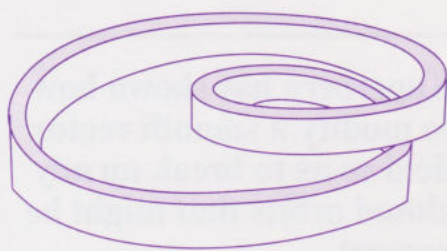
**Kuperberg has shown how to modify a smooth vector field so as to break up any closed orbits that might be present.**

---



**Figure 3.** The first stage in Kuperberg's counterexample to the Seifert conjecture is known as a Wilson plug. Trajectories that enter directly beneath the two circular orbits (dark lines) get trapped inside. (Figure courtesy of Krystyna Kuperberg.)

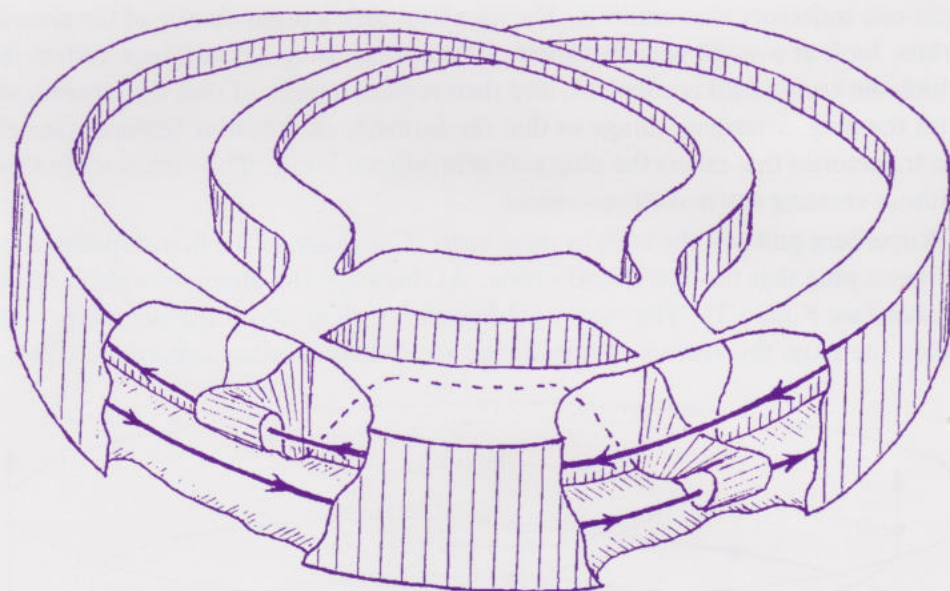




**Figure 4.** *The second stage in Kuperberg's construction.*

clockwise in the bottom half of the plug, but clockwise in the top half. As a result, any trajectory that makes it all the way through, exits the plug directly above where it enters—as though the field inside the plug were still constant. Trajectories that enter (and exit) near the inner and outer walls are deflected only slightly. The amount of deflection is greater for trajectories that enter away from the walls, until finally, trajectories that enter halfway between the walls pile up on a pair of circles pointing in opposite directions, and thus get trapped inside the plug.

In the second step, Kuperberg refashions the plug to make it look something like a pretzel (see Figure 4). The top and bottom surfaces of the plug no longer lie in a plane, but the walls remain vertical. Most important, there are now places where the inner and outer walls are close together. Finally, in the third step, Kuperberg pinches off two pieces of the plug near the outer wall and stuffs them through the inner wall, giving each piece a twist and skewering it on one of the closed trajectories (see Figure 5). These “self insertions” are the key to her counterexample.



**Figure 5.** *The final stage in Kuperberg's construction, with cut-away sections to show the self-insertions. (Figure courtesy of Krystyna Kuperberg.)*

“It’s the sort of thing that you wouldn’t think would be particularly helpful,” says Franks. To be sure, the self insertions break up the two closed orbits the plug began with. But it’s not at all clear that a slew of new closed orbits aren’t created in the process. Indeed, says Kuperberg, “if these self insertions are not chosen the right way, new closed trajectories may form.”

To control the trajectories created by the self insertions, Kuperberg relies on what she calls a “radius inequality.” Roughly speaking, when a self insertion satisfies this inequality, the resulting plug cannot contain any closed orbits. Instead, all the trapped trajectories spiral endlessly around inside the plug.

These endless trajectories do more than pull the plug on the Seifert conjecture. Kuperberg’s construction produces a “minimal set,” which John Mather, a dynamical systems theorist at Princeton University, suspects may be of an entirely new kind. Minimal sets are basic components of a dynamical system. Roughly speaking, a set is minimal if the dynamics on the whole set can be generated from



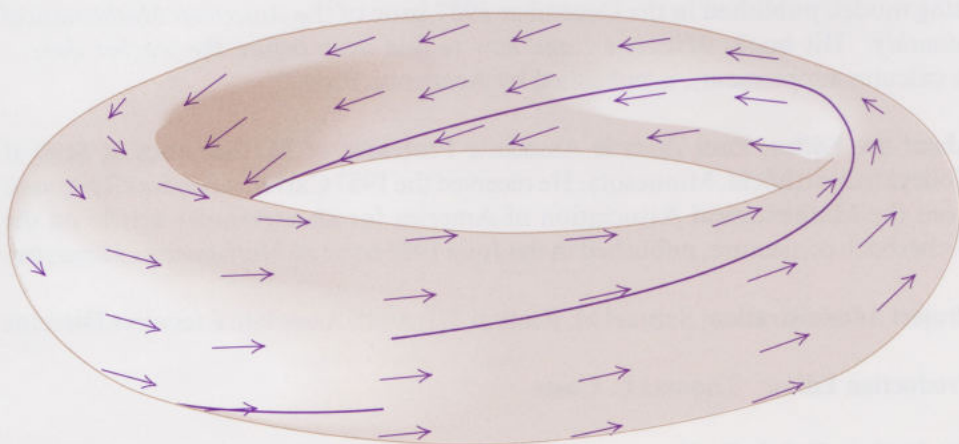
the dynamics on any piece of it. In particular, closed orbits are minimal sets, as are fixed points. Other kinds of minimal sets are known, says Mather, but “an overall picture of what minimal sets can be is just lacking.” By adding to the list of known examples, the minimal set contained in Kuperberg’s plug could help theorists better understand the range of things that can happen in dynamical systems.

Kuperberg’s construction also lowers the barrier to proving something much stronger, namely that the 3-sphere itself is minimal for the dynamical system associated with some vector field. Self-minimal manifolds are not that hard to find; the simplest example occurs on the torus (see box). Had the Seifert conjecture been true, the 3-sphere could not have been self-minimal, because a closed orbit can’t generate the dynamics away from itself. At this point there’s no hard evidence either way, but if it turns out the 3-sphere is self-minimal, that would do much more than refute the Seifert conjecture. It would darn near turn it inside out.

**By adding to the list of known examples, the minimal set contained in Kuperberg’s plug could help theorists better understand the range of things that can happen in dynamical systems.**

### Mathematical Donuts

Picture a chocolate cake donut with colored sprinkles all around it, all lined up to point in the same direction (see Figure 6). The non-caloric, mathematical version of this is known as a constant vector field on a torus. If you start at the outer circumference and follow the field once around, either you’ll advance along the circumference by a rational multiple of the circumference or you’ll advance by an irrational multiple. In the former case, the trajectory from any point will eventually close up; if the trajectory advances by the rational multiple  $m/n$ , it becomes periodic after  $n$  times around. But in the latter case, the trajectories never close up. Instead every trajectory winds about the torus forever, eventually coming arbitrarily close to every point on the surface. For such a vector field, the entire torus is a minimal set (see main story).



**Figure 6.** A vector field on a torus and part of one of its trajectories. The complete trajectory may or may not be closed. (Based on figure courtesy of Frederick Wicklin, Geometry Center, Minneapolis, Minnesota.)



---

## Credits

### ADVISORY BOARD

**Noga Alon**

Tel Aviv University

**Randolph E. Bank**

University of California, San Diego

**Robert Osserman**

Mathematical Sciences Research Institute

**Carl Pomerance**

University of Georgia

**Herbert S. Wilf**

University of Pennsylvania

---

**About the Author:** Barry Cipra, who also did the writing for volume 1 of *What's Happening in the Mathematical Sciences*, is a freelance mathematics writer based in Northfield, Minnesota. He is currently a Contributing Correspondent for *Science* magazine and also writes regularly for *SIAM News*, the newsletter of the Society for Industrial and Applied Mathematics. He received the 1991 Merten M. Hasse Prize from the Mathematical Association of America for an expository article on the Ising model, published in the December 1987 issue of the *American Mathematical Monthly*. His book, *Mistakes...and how to find them before the teacher does...* (a calculus supplement), is published by Academic Press.

**About the Editor:** Paul Zorn is Associate Professor of Mathematics at St. Olaf College in Northfield, Minnesota. He received the 1987 Carl B. Allendoerfer Award from the Mathematical Association of America for an expository article on the Bieberbach conjecture, published in the June 1986 issue of *Mathematics Magazine*.

**Project Administration:** Samuel M. Rankin, III, AMS Associate Executive Director

**Production Editor:** Thomas F. Costa

**Production:** Ralph E. Youngen, Neil G. Bartholomew, Lori E. Nero, Maxine Wolfson, and Lee Davol.

**Design:** Peter B. Sykes

The AMS gratefully acknowledges the support of the Alfred P. Sloan Foundation for the publication and distribution of *What's Happening in the Mathematical Sciences*.

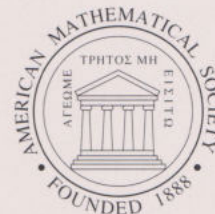


## About the American Mathematical Society

The American Mathematical Society is a non-profit organization devoted to research in the mathematical sciences. For more than 100 years, the Society has worked to support the advance of mathematical research, the communication of mathematical ideas, and the improvement of the mathematics profession. In recent years, the Society has increased its attention to mathematics education, public awareness of mathematics, and the connections of mathematics research to its uses.

The AMS is the world's largest mathematical organization, with nearly 30,000 members. As one of the world's major publishers of mathematical literature, the Society produces research journals and a wide range of books from advanced undergraduate texts to research monographs, as well as the authoritative reviewing and indexing journal: *Mathematical Reviews*. The Society is a world leader in the use of computer technology in publishing and is involved in the development of electronic means of information delivery. Another primary activity of the AMS is organizing meetings and conferences. In addition to an annual winter meeting, the Society organizes numerous smaller meetings during the academic year and workshops, symposia, and institutes during the summer. Other major Society activities are employment services, collection of data about the mathematical community, and advocacy for the discipline and the profession.

The main headquarters of the Society, located in Providence, Rhode Island, employs nearly 200 people and contains a large computer system, a full publication facility, and a warehouse. Approximately eighty people are employed at the *Mathematical Reviews* office in Ann Arbor, Michigan. The AMS also has an office in Washington, DC, in order to enhance the Society's public awareness efforts and its linkages with federal science policy.



To order *What's Happening in the Mathematical Sciences*:

**For orders with remittances:**

**(Payment must be made in U.S. currency drawn on a U.S. bank.)**

American Mathematical Society  
P. O. Box 5904  
Boston, MA 02206-5904, USA

**For credit card orders:**

American Mathematical Society  
P. O. Box 6248  
Providence, RI 02940-6248, USA  
1-800-321-4AMS (4267)  
1-401-455-4000, worldwide  
fax: 1-401-455-4046  
cust-serv@ams.org

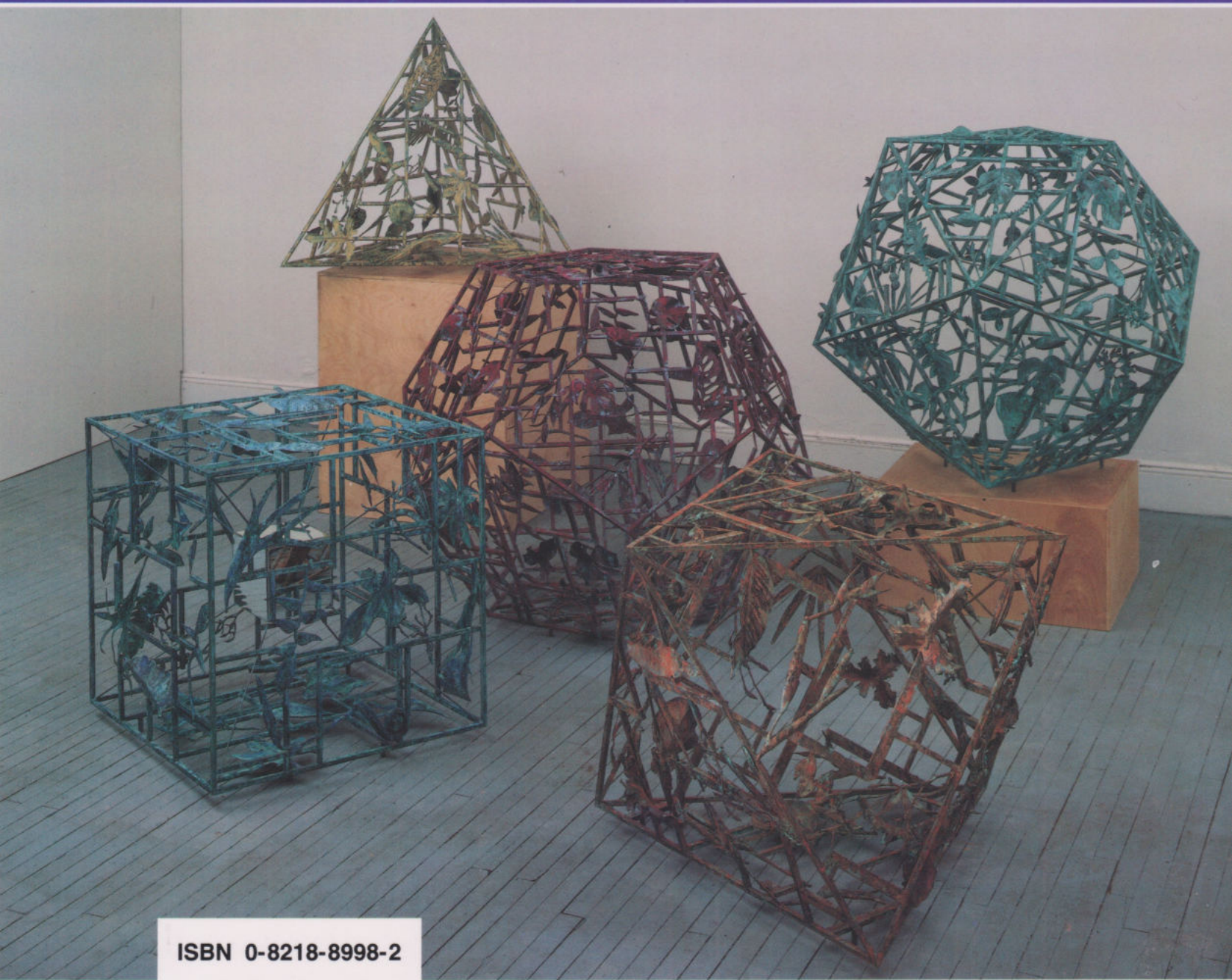
**For bookstore orders:**

American Mathematical Society  
P. O. Box 6248  
Providence, RI 02940-6248, USA  
or fax purchase orders to  
Sales Department,  
1-401-455-4046  
sales@ams.org

Volume 4; 1999; 126 pages; Softcover; ISBN 0-8218-0766-8; List **\$14**; Order code HAPPENING/4WH  
Volume 3; 1996; 111 pages; Softcover; ISBN 0-8218-0355-7; List **\$14**; Order code HAPPENING/3WH  
Volume 2; 1994; 51 pages; Softcover; ISBN 0-8218-8998-2; List **\$9.95**; Order code HAPPENING/2WH  
Volume 1; 1993; 47 pages; Softcover; ISBN 0-8218-8999-0; List **\$7**; Order code HAPPENING/1WH

All prices are subject to change. Charges for delivery are \$3.00 per order. For optional air delivery outside of the continental U.S., please include \$6.50 per item. Prepayment required. Or place your order through the AMS bookstore at [www.ams.org/bookstore/](http://www.ams.org/bookstore/). Residents of Canada, please include 7% GST.





ISBN 0-8218-8998-2



9 780821 889985